Kansalliskirjasto
Raportteja ja selvityksiä 4 / 2014

SUKUKIELTEN DIGITOINTIPROJEKTI

Pilottihankkeen loppuraportti

Jussi-Pekka Hakkarainen

Kansalliskirjasto, Helsinki 2014

ISBN 978-951-51-0056-6(pdf)
2. versio

KUVAILUSIVU

Julkaisija	Kansalliskirjasto
Julkaisun päivämäärä	20. 8.2014
Tekijä(t)	Jussi-Pekka Hakkarainen
Julkaisun nimi	Sukukielten digitointiprojekti Pilottihankkeen loppuraportti
Julkaisun osat	
Sarjan nimi ja numero	Raportteja ja selvityksiä 4/2014
ISSN	
ISBN	ISBN 978-951-51-0056-6
URL	
URN	URN:ISBN:978-951-51-0056-6 http://urn.fi/URN:ISBN:978-951-51-0056-6
Kokonaissivumäärä	49
Kieli	Suomi
Avainsanat – YSA (Suomi)	kielisukulaisuus uralilaiset kielet: http://finto.fi/yso/fi/page/p7464 kielentutkimus: http://finto.fi/yso/fi/page/p21012 kieliteknologia: http://finto.fi/yso/fi/page/p6071 tieteellinen yhteistyö: http://finto.fi/yso/fi/page/p10380 uhanalaiset kielet: http://finto.fi/yso/fi/page/p555 Sukukielten digitointiprojekti Digitization Project of Kindred Languages Проект по оцифровке родственных языков
Tiivistelmä	Kansalliskirjasto toteutti vuosina 2012–2013 Sukukielten digitointiprojektin pilottivaiheen. Pilottivaiheen loppuraportti esittelee hankkeelle asetettuja tavoitteita ja analysoi niiden saavuttamista. Loppuraportissa käydään läpi myös projektissa hyödynnettyjä tuotantomenetelmiä ja -välineitä, tarkastelee yhteistyötä niin venäläisten kirjastoalan partnerien kuin fennougristisen tutkimuksen kanssa, kuvaa hankkeen viestintää ja vertaa hanketta sen rahoittajan Koneen Säätiön Kieliohjelmaan.
Lisätietoja	

BESKRIVNING SIDA

Utgivare	Nationalbiblioteket
Utgivningsdatum	20.8.2014
Författare	Jussi-Pekka Hakkarainen
Publication	Digitalisering av material på finsk-ugriska släktspråk Pilotprojektets slutrapport
Publikationens delar	
Seriens namn och nummer	Rapporter och utredningar 4/2014
ISSN	
ISBN	ISBN 978-951-51-0056-6
URL	
URN	URN:ISBN:978-951-51-0056-6 http://urn.fi/URN:ISBN:978-951-51-0056-6
Sidoantal	49
Språk	Finska
Nyckelord (Allärs)	språkfrändskap uraliska språk: http://finto.fi/yso/fi/page/p7464 språkforskning: http://finto.fi/yso/fi/page/p21012 språkteknologi: http://finto.fi/yso/fi/page/p6071 vetenskapligt samarbete: http://finto.fi/yso/fi/page/p10380 hotade språk: http://finto.fi/yso/fi/page/p555 Sukukielten digitointiprojekti Digitization Project of Kindred Languages Проект по оцифровке родственных языков
Sammandrag	Åren 2012–2013 genomförde Nationalbiblioteket pilotstadiet i projektet för digitalisering av material på finsk-ugriska släktspråk. Slutrapporten för pilotstadiet presenterar projektets målsättningar och analyserar måluppfyllelsen. Rapporten presenterar också projektets produktionsmetoder och verktyg, samarbetet med aktörer i den ryska biblioteksektorn och med forskningen i fennougristik samt projektkommunikationen. Projektet har finansierats av stiftelsen Koneen Säätiö och i rapporten analyseras projektet i förhållande till Koneen Säätiös språkprogram.
Tilläggsuppgifter	

Sisältö

1 TIIVISTELMÄ 9

2 TAUSTAA 11

3 HANKKEEN KULKU 12

3.1 Projektiorganisaatio 12

3.1.1 Projektin työryhmä 12

3.1.2 Projektin ohjausryhmä 13

3.2 Tutkijayhteistyö 14

3.2.1 Aineistojen valinta osana yhteistyötä 14

3.2.2 Editorin kehittäminen tutkijoiden tarpeisiin 15

3.3 Yhteistyö venäläisten toimijoiden kanssa 16

3.3.1 Sopimukset 16

3.3.2 Tekijänoikeudet 16

3.3.3 Digitointi ja jälkikäsittely 17

4 PILOTTIVAIHEEN RISKIANALYYSI 19

4.1 Sopimukset 19

4.2 Venäläiset organisaatiot 21

4.3 Budjetti 22

4.4 Aikataulut 24

5 JULKAISUJÄRJESTELMÄ 26

5.1 Fenno-Ugrica –kokoelma julkaisujärjestelmänä 27

5.2 Fenno-Ugrican metadata 27

6 OCR-EDITORI KIELENTUTKIMUKSEN TUEKSI 29

6.1 Yleiskatsaus tekniseen ratkaisuun 29

6.2 Aineiston vastaanotto 29

6.3 OCR-editori 30

6.4 OCR-editorilla työskentely 30

6.5 OCR-editorin hallinnointi 31

6.6 Aineistojen tuominen OCR-editorista 32

7 VIESTINTÄ 33
7.1 Verkkosivut 33
7.2 Kansalliskirjasto-lehti 33
7.3 Blogi 33
7.4 Esitelmät 34
7.5 Sosiaalinen media 35
7.6 Kick off –tilaisuus 35

8 BUDJETTI 36

9 TULOKSET JA YHTEENVETO 39
9.1 Asetettujen tavoitteiden saavuttaminen 39
9.2 Pilotoinnin aikana opittua 40
9.3 Tulokset suhteessa Koneen Säätiön Kieliohjelmaan 42

10 SUOSITUKSET JATKOHANKKEEN TOTEUTTAMISTA VARTEN 44
10.1 Toteutustapa 44
10.2 Jälkikäsittely 45
10.3 Tutkijayhteistyö ja kansalaistieteet 46

11 YHTEYSTIEDOT JA TIEDUSTELUT 49

1 Tiivistelmä

Kansalliskirjasto toteutti Koneen Säätiön maaliskuussa 2012 myöntämällä apurahalla Sukukielten digitointiprojektin pilottihankkeen ajanjaksolla heinäkuu 2012-lokakuu 2013. Koneen Säätiön Kieliohjelmassa tähän osahankkeeseen viitataan termillä Pietarin aineistojen digitoinnin pilottihanke (2012–2013). Pilotoinnin tarkoituksena oli selvittää 1) edellytyksiä Venäjän Kansalliskirjaston [Российская национальная библиотека, Pietari] kokoelmissa olevien suomalais-ugrilaisia aineistojen digitaalisesta käyttöön saattamista, 2) aineistoihin kohdistuvien tekijänoikeuksia ja 3) aineistojen soveltuvuutta tutkimuskäytössä

Pilottihankkeessa digitoitu aineisto sisälsi 17000 sivua marin-, mordvan-, inkeroisten- ja vepsänkielisiä julkaisuja, yhteensä 156 monografiaa, jotka ovat pääsääntöisesti neuvostoajan alun oppikirjoja ja sanakirjoja. Monografioiden lisäksi digitoitiin lähes 25 000 sivua marilaisia ja mordvalaisia sanomalehtiä lähinnä 1920- ja 30-luvuilta. Ensimmäiset pilottihankkeen digitoidut aineistot asetettiin avoimeen käyttöön kesäkuussa 2013 ja viimeiset materiaalit kuvailtiin Fenno-Ugrica -kokoelmaan [http://fennougrica.kansalliskirjasto.fi/] syyskuussa 2013.

Pilottiprojektisssa on työskennelty niin kansainvälisen kirjastokentän kuin kotimaisen kielentutkijayhteisön kanssa. Kansainvälistä työnjakoa on ohjannut lähtökohtaisesti Suomen ja Venäjän Kansalliskirjaston välinen yhteistyösopimus, johon nojaten yhteistyö voitiin käynnistää. Hanketta varten neuvoteltiin syksyllä 2012 kirjastojen välinen sopimus, missä määriteltiin hankkeen puitteissa toteutettavat kummankin osapuolen vastuut ja velvollisuudet digitoitavien aineistojen suhteen.

Aineistojen tekijänoikeuksien selvittämisestä ja korvaamisesta sovittiin erikseen Kansalliskirjaston ja venäläisen kirjastoalan tekijänoikeuksia selvittävän toimiston National Library Resourcen (venäjäksi Национальный библиотечный ресурс, Moskova) välillä, joka suoritti digitoitaviin aineistoihin kohdistuvan tekijänoikeuksienselvityksen joulukuun 2012 ja maaliskuun 2013 välisenä aikana. Selvitykseen nojautuen, Kansalliskirjasto päätti julkaista digitoidun aineiston avoimella palvelimella, jotta projekti vastaisi Koneen Säätiössä Kieliohjelmassa esitettyjä vaatimuksia aineistojen saatavuudesta, saavutettavuudesta ja käytettävyydestä. Kansainvälisesti merkittävän hankkeesta on tehnyt se, että molemmat hankkeessa mukana olleet venäläiset partnerit olivat ensimmäistä kertaa toteuttamassa tämänsuuntaista digitointihanketta ulkomaalaisten yhteistyökumppanin kanssa. Digitointihanke selvitystöineen on kansainvälisesti ainutlaatuinen – koskaan aikaisemmin ei venäläisten kirjastojen kokoelmista digitoituja aineistoja ole julkaistu toisen maan alueella sijaitsevan kirjaston kokoelmissa tässä laajuudessa ja tekijänoikeudet huomioiden.

Kotimaiseen suomalais-ugrilaisia kieliä tutkivaan tiedeyhteisöön oltiin yhteydessä pilotoinnin aikana tiiviisti. Yhteistyö oli luonteeltaan keskustelevaa ja tutkijalähtöistä, ja hankkeessa pyrittiin ottamaan huomioon tutkijoiden mielipiteet ja toiveet mahdollisimman laajasti. Kieliohjelmassa esitettyihin vaatimuksiin pyrittiin vastaamaan myös luomalla tutkimusinfrastruktuuria, jonka avulla tutkijakunta pystyy työstämään digitoitua aineistoa omaa tutkimustaan varten. Yksi näistä tutkimusinfrastruktuurin työkaluista on pilottihankkeen aikana kehitelty OCR-editori, jonka avulla aineistojen sähköiseen muotoon tallennettua tekstiä voidaan muokata. Työkalun kehittämisessä Kansalliskirjasto ja tutkijakunta toimivat yhteistyössä. Digitoitavien aineistojen valinnassa ja niiden evaluoinnissa tutkijakunnalla oli keskeinen rooli.

Erityisesti suomalais-ugrilaisten kielten alkuperäispuhujille aineistot ovat tarjonneet monisyisen mahdollisuuden tarkastella menneisyyttään, kun taas muille loppukäyttäjille Fenno-Ugricassa julkaistut aineistot tarjoavat uuden mahdollisuuden tutustua suomalais-ugrilaisten kansojen kieliin ja kulttuureihin. Pilottihankkeen aikana oli tarkoitus saada avoimeen käyttöön sellaisia aineistoja, joita ei ole saatavilla toisaalla digitaalisessa muodossa ja etenkään tässä laajuudessa. Vaikka digitoitavan aineiston määrä tuntuu kovin pieneltä, niin kyseessä on tarkoin harkittuja kokonaisuuksia, jotka palvelevat laajempia tarkoituksia, ja kuten Koneen Säätiön Kieliohjelmassa sanotaan, ”tarkoituksena on pienten suomalais-ugrilaisten kielten, suomen sekä Suomen vähemmistökielten dokumentointi ja niiden aseman vahvistaminen.” Tämän edellytyksenä on ollut digitoivan aineiston mahdollisimman vapaa saavutettavuus ja tavoitteena olikin, että pilottihankkeen aineisto saatettaisiin avoimeen käyttöön myös tulevan Kansallisen digitaalisen kirjaston asiakasliittymän (Finna) ja vastaavien kansainvälisten palveluiden (mm. Europeana) avulla. Aineistot ovat löydettävissä myös Googlella ja muilla yleisillä hakukoneilla. Lisäksi aineistoa voidaan etsiä Kansalliskirjaston ylläpitämästä Uralica-portaalista [uralica.kansalliskirjasto.fi], johon on linkitetty uralilaisilla kielillä digitoituja kieliaineistoja eri kirjastojen digitaalisista kokoelmista.

Sukukielten digitointiprojektin pilottihankkeessa on onnistuttu saavuttamaan kiitettävällä tavalla ne tavoitteet, joita pilottiprojektille oli asetettu. Aineistojen käyttöön saattamisen kannalta keskeiset tekijänoikeudelliset kysymykset ratkaistiin yhteistyössä venäläisten toimijoiden kanssa ja tämän saavuttaminen oli keskeinen edellytys hankkeen onnistumisen kannalta. Pilotoinnin aikana on todettu, että Venäjän Kansalliskirjaston kokoelmista voidaan digitoida ja asettaa avoimeen käyttöön sellaista aineistoa, joka on tutkimuksellisesti merkittävää. Tähän saavutukseen on päästy uupumatomalla yhteistyöllä venäläisten partnerien kanssa.

2 Taustaa

Koneen Säätiö käynnisti vuonna 2012 kieliohjelman, jonka yhtenä tavoitteena on dokumentoida pieniä suomalais-ugrilaisia kieliä, suomea ja Suomen vähemmistökieliä. Kieliohjelman osana Koneen Säätiö oli syksyllä 2011 käynnistämässä digitointihanketta, jossa tarkoituksena oli digitoida Venäjällä olevia suomalais-ugrilaisia julkaisuja. Tuolloin Koneen Säätiö tiedusteli Kansalliskirjaston ja Helsingin yliopiston kirjaston halukkuutta toimia digitointihankkeen suomalaisena tilaajana. Keskinäisten neuvotteluiden jälkeen Kansalliskirjasto päätti ensin selvittää, olisiko hanke ylipäänsä mahdollista toteuttaa ja millä ehdoilla.

Alkuvuodesta 2012 Jussi-Pekka Hakkarainen selvitti Kansalliskirjaston toimeksiannosta edellytyksiä toteuttaa nk. Pietarin aineistojen digitoinnin pilottihanke (2012–2013), jonka tarkoituksena oli kehittää tuotantoprosessi ja luoda käytännöt eri osapuolten yhteistyönä toteutettavalle aineiston digitoinnille, käyttöön asettamiselle ja aineiston säilytykselle. Tavoitteena oli, että pilottihankkeen tulokset olisivat hyödynnettävissä suomalais-ugrilaisen kielikorpuksen kehittämisessä, siihen liittyvissä tutkimushankkeissa ja että aineistot ovat vapaasti saatavilla internetissä.

Selvityksestä syntyneessä projektisuunnitelmassa hahmoteltiin se infrastruktuuri, missä edellä mainittu suomalais-ugrilaisten aineistojen pilottihanke voitaisiin toteuttaa. Selvitystyön aikana keskusteltiin niin Koneen Säätiön, tutkijayhteisön edustajien, Kansalliskirjaston, Helsingin yliopiston kirjaston kuin Venäjän Kansalliskirjaston (Pietari) edustajien kanssa hankkeen toteuttamisesta yksityiskohtaisesti.

Selvityksessä hahmoteltiin digitoinnin, digitoidun aineiston käyttöön asettamisen, julkaisujärjestelmän ylläpidon ja viestinnän prosessit. Selvityksessä kuvattiin pilottihankkeen tekninen toteuttamistapa ja määriteltiin digitoitua aineistoa koskevat vaatimukset, myös aineiston metadataa koskien. Selvityksessä tarkasteltiin ja esiteltiin digitoitavaan aineistoon liittyviä tekijänoikeuksia ja niihin liittyviä kysymyksiä. Lisäksi selvityksessä otettiin kantaa yhteistyöosapuolien keskinäiseen työnjakoon ja annettiin ehdotus hankkeen organisaatiosta (ohjausryhmä, taloushallinto ja viestintä) sekä sen järjestämisestä.

Selvityksen pohjalta muotoillulla projektisuunnitelmalla Kansalliskirjasto haki maaliskuussa 2012 Koneen Säätiöltä apurahan pilottihankkeen toteuttamiseksi vuosina 2012–2013. Apuraha myönnettiin Kansalliskirjastolle haetun, eli 356 040 euron suuruisena

3 Hankkeen kulku

Sukukielten pilottiprojekti käynnistyi heinäkuussa 2012 ja loppui lokakuussa 2013. Hankkeen kulkua ja sen toteuttamista käsitellään tässä raportissa kolmen toiminta-ympäristön kautta. Ensimmäinen niistä liittyy projektin paikalliseen toteuttamiseen Kansalliskirjastossa. Toisen kehikon sisällä tarkastellaan venäläisten partnerien kanssa harjoitettua yhteistyötä, kun taas kolmas osa-alue tarkastelee projektin aikana harjoitettua tutkijayhteistyötä.

3.1 Projektiorganisaatio

Sukukielten digitointiprojektilla ei ole ollut varsinaista omaa, muusta Kansalliskirjaston toiminnasta erillistä työyksikköä ja hankkeen projektipäällikkö on työskennellyt projektille ilman Kansalliskirjaston yleiseen toimintaan tai muihin projekteihin liittyviä velvoitteita.

3.1.1 Projektin työryhmä

Hankkeelle koottiin elokuussa 2012 työryhmä, johon kuuluivat projektipäällikön lisäksi kirjastonhoitaja Marina Ivanova (Slaavilainen kirjasto), tietojärjestelmäasiantuntija Juho Vuori (Kirjastoverkkopalvelut) ja taloussuunnittelija Marja-Liisa Lonardi (Hallintopalvelut). Ivanovan ja Vuoren työpanos korvattiin heidän omalleen työyksikölle tuntikirjanpidon mukaisesti projektin varoista. Joulukuussa 2012 hankkeeseen palkattiin tietojärjestelmäasiantuntijaksi Wouter van Hemel, jonka palkkaukseen käytettiin sekä Sukukielten digitointiprojektin (6kk) että Kirjallisuuspankkihankkeen (9kk) varoja. Lisäksi hankkeeseen palkattiin kirjastosihteeriksi Viktoria Kurkina ajalle heinä-syyskuu 2013.

Lonardin työtehtävät hankkeessa keskittyivät projektipäällikön tukemiseen taloushallinnon, erityisesti budjetointiin liittyvissä tehtävissä. Vuori osallistui Helsingin yliopiston asettaman rekrytointikiellon ajan OCR-editorin suunnitteluun ja van Hemel jatkoi Vuoren työtä joulukuusta 2012 lähtien. Ivanovan ja Kurkinan työtehtävät keskittyvät digitoidun aineiston kuvailuun ja tarvittaessa käännöstehtäviin. Lisäksi Ivanova teki paljon asiantuntijatyötä liittyen kuvailuformaatteihin. Projektipäällikkönä toimineen Jussi-Pekka Hakkaraisen tehtäväkenttä oli laaja. Hänen työtehtäviinsä kuuluivat mm. yhteistyön koordinointi kotimaisten ja ulkomaisten toimijoiden kanssa, projektin viestintään, raportointiin, Fenno-Ugrica -kokoelmaan ja taloushallintoon liittyvät työtehtävät. Projektityöryhmä kokoontui useimmiten

hankkeen alkuvaiheissa, ennen kaikkea syksyn 2012 ja alkutalven 2013 aikana. Kun työryhmän jäsenten työtehtävät eriytyivät projektin edetessä, yhteiset kokoontumiset lopetettiin ja projektipäällikkö tapasi työryhmän jäseniä lähes viikottain heidän työtehtäviensä tiimoilta.

Työryhmän jäsenten lisäksi projektissa ovat eri tavoin auttaneet muu Kansalliskirjaston henkilökunta. Kansalliskirjaston Digitointi- ja konservointikeskuksen (Mikkeli) henkilökunta on auttanut hanketta erityisesti OCR-editorin XML-formaatin hallintaan liittyvissä kysymyksissä. Kansalliskirjaston Kirjastoverkkopalvelut ovat tarjonneet asiantuntemustaan projektin käyttöön erilaisissa teknisissä ja metadatan hallintaan liittyvissä työvaiheissa. Kansalliskirjaston Tutkimuskirjastopalveluiden henkilökunta on auttanut projektipäällikköä liittyen yhteistyöhön sekä tutkijakunnan kuin ulkomaisten toimijoiden kanssa.

3.1.2 Projektin ohjausryhmä

Projektille kutsuttiin ohjausryhmä, jossa oli edustettuna niin Kansalliskirjasto, Helsingin yliopiston kirjaston, rahoittajan kuin tutkijoiden edustajia. Ohjausryhmää valittaessa ja sitä koolle kutsuttaessa Sukukielten digitointiprojekti ja Kansalliskirjasto halusivat, että mukana olisi mahdollisimman laaja-alainen osaajien joukko, jotka pystyisivät ohjaamaan hanketta useasta eri asiantuntijaperspektiivistä ja että ohjausryhmän tapaamisen toimisivat myös informaatiokanavana rahoittajan suuntaan.

Ohjausryhmän jäsenet olivat:

- *Helsingin yliopisto*
  Pauli Assinen
- *Koneen Säätiö*
  Risto Alapuro
- *Tutkijakunnan edustajat*
  Jack Rueter, Helsingin yliopisto
  Jorma Luutonen, Turun yliopisto
- *Kansalliskirjasto*
  Liisa Savolainen, johtaja, projektin omistaja, ohjausryhmän puheenjohtaja
  Leena Saarinen, tuotantopäällikkö
  Jussi-Pekka Hakkarainen, projektipäällikkö, esittelijä
  Elli Salo, ohjausryhmän sihteeri

Ohjausryhmä kokoontui tässä kokoonpanossa yhteensä viisi kertaa syyskuun 2012 ja marraskuun 2013 välisenä aikana. Ohjausryhmän vakinaisten jäsenten lisäksi kokouksiin kutsuttiin vierailevia asiantuntijoita tarpeen mukaan.

Pilottiprojektin etenemistä on seurattu ja ohjattu Kansalliskirjaston sisällä kahdella eri taholla: projektipäällikkö on esitellyt hanketta säännöllisesti Tutkimuskirjaston johtoryhmässä vähintään kerran kuukaudessa. Tämän lisäksi projektipäällikkö on informoinut Kansalliskirjaston johtoryhmää kahteen otteeseen hankkeen etenemisestä.

3.2 Tutkijayhteistyö

Sukukielten digitointiprojektissa on pyritty rakentamaan tiivistä ja osallistuvaa yhteistyötä tutkijakunnan ja Kansalliskirjaston välille. Hanke on alusta lähtien ollut voimakkaasti tutkijalähtöinen ja se on tähdännyt tutkijakunnan kokonaisvaltaiseen palvelemiseen. Käytännön tasolla tämä on tarkoittanut sitä, että tutkijoita on sitoutettu mukaan hankkeen suunnitteluun, toteutukseen ja päätöksentekoon.

Hanketta suunnitellessa tutkijoiden aloitteet ja verkostot ovat olleet ratkaisevia. Kansalliskirjasto ei aikaisemmin ole toiminut näin voimakkaasti suomalais-ugrilaisen maailman ja kielentutkimuksen kentällä, joten hankkeen toteuttajaa voidaan pitää siinä mielessä uutena toimijana tällä kentällä. Tässä suhteessa tutkijakunnan kontaktit ja näkemykset ovat olleet hankkeen toteuttamisen kannalta oivallisia voimavaroja, joihin Sukukielten digitointiprojektissa on pyritty parhaan mukaan tutustumaan.

Tutkijoiden kutsuminen mukaan hankkeen ohjausryhmään ja lähemmäksi päätöksentekoa on ollut myös hankkeen toteuttamisen kannalta hedelmällistä ja tutkijakunnan edustajat ovat voineet täten vaikuttaa projektin etenemiseen heidän omista premisseistään lähtien. Tutkijoiden läsnäolo on lisännyt osapuolten keskinäistä ymmärrystä eri toimintakulttuureista. Seuraavassa esitellään kaksi toiminnan kenttää, missä tutkijayhteistyöllä on ollut ratkaiseva osuus.

3.2.1 Aineistojen valinta osana yhteistyötä

Aineistojen valinta ja valinnan kriteerien määrittely ei olisi ollut mahdollista ilman tutkijoiden osallistumista hankkeeseen. Tutkijat tuntevat parhaiten aineistoihin kohdistuvat tarpeet ja odotukset, joten tutkijoiden äänen kuuluminen oli tässä kohtaa ensisijaista.

Keskeisimpänä aineiston valintakriteerinä on ollut nykykirjakielten synty- ja vakiintumisajankohta. Komin, udmurtin, vuori- ja niittymarin, inkeroisen, ersän, mokšan ja vepsän osalta tämä prosessi tapahtui maailmansotien välisenä aikana, keskeisimmin vuosina 1930-luvun alkupuolella.

Sukukielten nykyisille vaalijoille kirjakielen syntykauden aineisto on tärkeää: 1920- ja 1930-luvun uudissanat ja niitä käyttävät tekstit toimivat meidän aikamme kielenkehittäjille yhtä lailla lähdemateriaalina kuin innovaatioiden ja innoituksen lähteenä. Digitoidut aineistot on valittu siten, että ne yhtäältä kuvastavat mahdollisimman hyvin 1920-luvun innovatiivista aikaa mutta toisaalta ilmentävät myös sitä kielipoliittista muutosta, joka tapahtui 1930-luvulla. Tällaisten aineistoihin kohdistuvien määrittelyjen tekeminen ilman tutkijoiden osallistumista ja heidän asiantuntemusta ei olisi ollut mahdollista.

Kieliaineistojen valintoja rajoittavia tekijöitä on ollut muutamia. Yksi keskeinen rajoitus on ollut aineiston saavutettavuus Suomesta. Mikäli teos löytyy joko Kansalliskirjaston tai suomalaisten yliopistonkirjastojen kokoelmista, teosta ei ole otettu mukaan digitoitavaksi, sillä tutkijoiden on ajateltu pääsevän käsiksi aineistoon muutoinkin.

Toinen merkittävä rajaava tekijä on ollut aineistoihin kohdistuvat tekijänoikeudet. Ennen vuotta 1941 julkaistua aineistoa voidaan saattaa avoimeen tutkija- ja kansalaiskäyttöön verrattain helpostikin, mutta toisen maailmansodan aikainen ja sen jälkeinen aineisto edellyttää väistämättömästi sopimista tekijänoikeudenhaltijoiden kanssa. Tietoisuus tekijänoikeuksista ja niihin liittyvistä ongelmista on ohjannut aineistojen valintaa ja siten painostus on vahvasti 1920- ja 30-luvuilla julkaistuissa aineistoissa.

3.2.2 Editorin kehittäminen tutkijoiden tarpeisiin

Pilottiprojektissa toteutettiin web-käyttöliittymä digitoitujen kirjojen ja sanomalehtien kuvista tunnistetun tekstin korjaamiseen ja muuhun editointiin. Tekstintunnistus (OCR) tuottaa virheellisiä merkkijonoja erityisesti huonolaatuisista originaaleista, sekä kielistä, joiden korpuksia ei vielä ole saatavilla. Virheet on korjattava, että tunnistettua tekstiä voi jatkokäyttää.

Yhteistyö tutkijoiden kanssa voidaan sanoa olleen hankkeen onnistumisen kannalta keskeisellä sijalla ja tutkijat ovat olleet mukana edistämässä OCR-editorin suunnittelua eri kehitysvaiheissa. Ilman heidän kontribuutioitaan, OCR-editoriin liitetyt toiminnallisuudet eivät välttämättä vastaisi niitä tavoitteita, mitä sille asetettiin ja millaiseen työhön tutkijat editoria käyttäjät. Muutamia sellaisia toiminnallisuuksia, joihin aloite on tullut tutkijoilta, ovat mm.:

- virtuaalinäppäimistö, jossa näytetään teoksen kielen merkit
- ylläpidon ja peruskäyttäjän erottaminen
- automaattisten muutosten ilmoittaminen
- korpuksen näyttö
- linkit sivuihin, joilla editoitava sana tai sanamuoto on.
- sanojen järjestäminen aakkosjärjestykseen ja lukumäärän mukaan.

3.3 Yhteistyö venäläisten toimijoiden kanssa

Sukukielten digitointiprojektin yksi isoimmista haasteista oli yhteistyö venäläisten toimijoiden kanssa. Pilottivaiheen aikana Kansalliskirjasto harjoitti tiivistä yhteistyötä ennen kaikkea moskovalaisten tekijänoikeusjärjestö National Library Resourcen kanssa, joka selvitti aineistoihin koskevia tekijänoikeuksia ja Venäjän Kansalliskirjaston kanssa, joka digitoi ja jälkikäsitteli pilottihankkeessa tuotettavan aineiston. Molemmat partnerit olivat hankkeen aikana myös sopimusosapuolia. Edellä mainittujen toimijoiden lisäksi Kansalliskirjasto pyrki keskustelemaan Marin, Karjalan, Komin, Mordovian ja Udmurtian tasavaltojen kansalliskirjastojen kanssa yhteistyöstä mahdollisen jatkohankkeen aikana.

3.3.1 Sopimukset

Sukukielten digitointiprojektin pilottivaiheessa Kansalliskirjasto teki uusia sopimuksia venäläisten osapuolten kanssa. Sopimuksissa määriteltiin osapuolten vastuut ja velvollisuudet, ostopalveluiden hinnat ja aikataulut.

Syksyllä 2012 Kansalliskirjasto sopi aiesopimuksen National Library Resourcen kanssa mahdollisesti tekijänoikeuksien korvaamisesta ja tekijänoikeuksien selvitystyön kustannuksista. Kun National Library Resource sai maaliskuussa tekijänoikeuksien selvityksen valmiiksi, niin tuolloin ei ollut ilmennyt ainoatakaan tekijänoikeudenhaltijaa, jonka kanssa olisi voitu tekijänoikeuksien siirrosta ja niiden korvaamisesta sopia. Tekijänoikeuksista ei siis ole sovittu, vaan ainoastaan niiden selvittämisestä on maksettu korvaus National Library Resourcelle.

Venäjän Kansalliskirjaston kanssa sovittiin erikseen aineiston digitoinnista, sen jälkikäsittelystä ja näiden palveluiden hinnoista. Sopimus sisälsi aikataulun, osapuolten vastuut ja velvollisuudet sekä käytännöt mahdollisten kiistatilanteiden ratkaisemiseksi välimiesoikeudessa.

3.3.2 Tekijänoikeudet

Pilottiprojektin onnistumisen kannalta keskeisimmät kysymykset liittyivät venäläiseen tekijänoikeuslainsäädäntöön ja sen soveltamiseen Sukukielten digitointiprojektissa. Keskeinen periaate on se, ettei aineistoa voida digitoida ilman tekijänoikeudenhaltijoiden suostumusta. Sukukielten pilottiprojektissa tekijänoikeudenhaltijoiden suostumus aineistoa koskien haluttiin sopia mahdollisimman laajasti, jopa niin että aineistoa voitaisiin esittää rajoituksetta tietoverkoissa ja sitä voitaisiin jakaa edelleen kolmansille osapuolille.

Kansalliskirjastolta oltiin jo hankkeen suunnitteluvaiheessa yhteydessä National Library Resourceen, moskovalaiseen tekijänoikeusjärjestöön, jonka kanssa tuli

keskustella ja mahdollisesti sopia tekijänoikeuksien selvittämisestä. Tämä järjestely johtui siitä, että Venäjän Kansalliskirjasto oli antanut mandaattinsa tekijänoikeuksien selvittämisestä tälle organisaatiolle.

National Library Resourceen oli oltu jo talven 2012 aikana selvitystyön yhteydessä, mutta tuolloin heihin ei oltu saatu aitoa keskustelukontaktia, vaikka Venäjän Kansalliskirjastolta oli heitä pyydetty vastaamaan tiedusteluihimme. Tekijänoikeuksien selvittäminen oli hankkeen toteuttamisen kannalta pullonkaula, sillä ilman niiden selvittämistä aineistoa ei voisi julkaista avoimella palvelimella ja tällöin Koneen Säätiön Kieliohjelman tavoitteet avoimuudesta eivät kokonaisuudessaan toteutuisi.

Keskusteluyhteys saatiin aikaan syyskuussa 2012, jolloin National Library Resource lopulta vastasi projektipäällikön tiedusteluihin. Lokakuussa 2012 projektipäällikkö vieraili järjestön luona Moskovassa, missä sovittiin digitoitavan aineiston tekijänoikeuksien selvittämisestä. National Library Resource ehdotti, että he suorittavat selvityksen jokaista digitoitavaa nimekettä koskien ja mikäli tekijänoikeuden haltijoita ilmenisi, tulisi heidän kanssaan sopia vasta selvityksen päätyttyä. National Library Resourcen puolelta selvitystä koordinoi Aleksandr Balatski ja selvityksen lainopillisena neuvonantajana toimi professori ja lakimies Vladimir Entin Klishin & Partners -nimisestä tekijänoikeuksiin erikoistuneesta lakiasiaintoimistosta.

Kansalliskirjasto hyväksyi National Library Resourcen ehdotuksen selvitystyön suorittamisesta ja osapuolet (Kansalliskirjasto ja National Library Resource) allekirjoittivat sopimuksen selvitystyöstä joulukuun 2012 alussa, jolloin selvitystyö myös aloitettiin. Selvitystyön kestoksi määriteltiin 90 arkipäivää. Selvitystyön aikana National Library Resource yritti tavoittaa tekijänoikeudenhaltijoita lehti-ilmoituksin, mm. Literaturnaja gazetan vuoden 2013 ensimmäisessä numerossa (ilm. 16.tammikuuta 2013) ja Joshkar-Olassa ilmestyvässä Marii El -lehdessä 12. tammikuuta 2013. Lisäksi National Library Resource selvitti tekijänoikeudenhaltijoiden mahdollisia elinvuosia kahdeksalta eri viranomaistaholta ja arkistoilta.

Hankkeessa tehty selvitys oli poikkeuksellinen ja tällainen selvitys tekijänoikeuksienhaltijoista tässä laajuudessaan suoritettiin Venäjän nykyoloissa ensimmäistä kertaa. Koskaan aikaisemmin koti- tai ulkomainen kirjasto ei ollut selvittänyt aineistojen oikeuksia yhtä laajasti tavoitteenaan avoin aineiston käyttö.

3.3.3 Digitointi ja jälkikäsittely

Tilattujen aineistojen digitointi ja jälkikäsittely suoritettiin Venäjän Kansalliskirjastossa, jonka kokoelmiin teokset kuuluivat. Pilottihankkeen aikana digitoidun aineiston teknisiä vaatimuksia määriteltiin yhdessä Venäjän Kansalliskirjaston kanssa jo selvitysvaiheen aikana talvella 2012 ja OCR-tekstin osalta uudestaan syksyllä 2012.

Koska aineistoja ei voida luovuttaa Suomeen digitoitavaksi, digitointi suoritetaan Venäjän Kansalliskirjastossa ja heidän digitoimansa aineistot toimitettiin edelleen Kansalliskirjastolle. Monografiat digitoitiin harmaasävyskannattuna 300 dpi tarkkuuteen, kun taas sanomalehtien resoluutioksi määriteltiin 400 dpi. Primääriskannaus suoritettiin TIFF-kuvaformaattiin ja skannattujen kuvien yhteyteen lisättiin automaattisen tekstintunnistamiseen (OCR, optical character recognition) vaadittavat elementit. OCR-teksti liitettiin kuvatiedostoon ja nämä elementit tuotiin yhdeksi tiedostoksi, PDF-formaattiin, joka oli myös aineiston esitysformaatti julkaisujärjestelmässä. Mahdollista myöhempää käsittelyä varten myös TIFF-muotoiset kuvat luovutettiin Kansalliskirjastolle.

Varsinainen digitointityö onnistui Venäjän kansalliskirjastossa erinomaisesti ja kaikki hanketta varten digitoidut aineistot eräitä konservoitavia sanomalehtiä lukuun ottamatta olivat valmiina jo huhtikuun puolessa välissä. Aineistojen toimittaminen Helsinkiin kuitenkin viivästyi huomattavasti, sillä OCR-tekstin lisääminen vei Venäjän kansalliskirjastolta enemmän aikaa kuin oli arvioitu. Tässä työvaiheessa ongelmia ilmeni sekä paikallisen työn koordinoinnissa että sen resursoinnissa. Venäjän Kansalliskirjasto oli järjestänyt työprosessien sisäisen koordinoinnin Ulkomaisten asioiden osaston kautta, joka pyrki koordinoimaan hankkeessa tehtävää työtä kirjaston eri osastoilla. Tämä ei kuitenkaan riittänyt työn laadukkaaseen tekemiseen kaikissa työvaiheissa osastoilla, vaan OCR-tekstin lisääminen viivästytti pahasti aineiston toimittamista. Toukokesäkuussa tehtyjen tiedustelun perusteella ongelmaksi venäläisten puolelta ilmoitettiin ennen kaikkea työvaiheen aliresursointi. Viivästys koski ennen kaikkea sanomalehtiaineistoa, jonka viimeiset numerot saatiin kuvailluksi Fenno-Ugricaan vasta syyskuun 2013 lopussa, yli neljä kuukautta sopimuksessa olleen viimeisen toimituspäivän (9.toukokuuta 2013) jälkeen.

Yksi ongelma tämän työvaiheen laadukkaassa toteuttamisessa oli se, etteivät venäläiset kirjastot vielä ole tuottaneet paljoakaan OCR-tekstiä sisältäviä aineistoja. Tämä tuotannon tapa oli Venäjän Kansalliskirjastolle verrattain uusi ja kokemattomuus näkyi mm. siinä, että Sukukielten digitointiprojektin työntekijät kouluttivat paikallista henkilökuntaa käyttämään ohjelmistoa sopimuksen edellyttämällä tavalla. Toinen, sekä tähän työvaiheeseen että kirjastojen väliseen yhteistyöhön yleensäkin liittyvä aspekti on se, että suomalaisen ja venäläisen työkulttuurin erilaisuus aiheutti esteitä keskinäisessä viestinnässä. Hankkeen aikana yleisestikin huomattiin, että venäläisen työkulttuurin hierarkkisuus esti usein sikäläisten työntekijöiden omatoimisen ongelmanratkaisun.

Yleisesti Venäjän Kansalliskirjaston tuottamaan digitoitiin ja OCR-tekstin laatuun oltiin tyytyväisiä, mutta jatkohankkeessa OCR-tekstin tuottamista suositellaan poissiirrettäväksi Pietarista, jotta työvaihetta voidaan kontrolloida laadukkaammin kuin mitä pilottihankkeessa on ollut mahdollista, kasvavien kustannuksien uhallakin. Mikäli työvaihe siirretään Suomeen, niin OCR-tekstin laatua voidaan parantaa prosessinomaisesti yhteistyössä tutkijoiden kanssa. Tämä tuotantotavan muutos tarkoittaisi sitä, että jatkossa vastaanotettaisiin digitoinnit ainoastaan TIFF-muotoisina kuvina ja jälkikäsittelyt tehtäisiin Suomessa.

4 Pilottivaiheen riskianalyysi

Ohjausryhmän kokouksessa 10. lokakuulta 2012 keskusteltiin pilottivaiheen projektisuunnitelmasta ja sen toteuttamisesta. Ohjausryhmän jäsen, Pauli Assinen (HYK), huomautti, ettei projektisuunnitelma varsinaisesti sisältänyt riskianalyysiä. Ohjausryhmän pyynnöstä hankkeen projektipäällikkö valmisteli riskianalyysin, missä pyrittiin kartoittamaan ja paikantamaan mahdolliset projektin toteuttamiseen liittyvät riskit, arvioimaan niiden todennäköisyys ja esittämään miten näitä riskejä voidaan hallita tai poistaa kokonaan.

Riskien suuruuden arvioinnissa hyödynnettiin kuuden muuttujan kenttää. Riskitaulukossa seurausten vakavuudelle ja tapahtuman todennäköisyydelle on kolme eri tasoa. Tehtyjen selvitysten perusteella valitaan ensiksi seurausten vakavuus taulukon ylimmältä riviltä ja sen jälkeen tapahtuman todennäköisyys ensimmäisestä sarakkeesta. Riski on valittujen kohtien leikkauspisteessä olevan arvon suuruinen. Riskin suuruus saa pienimmillään arvon 1 (Merkityksetön riski) ja suurimmillaan arvon 5 (Sietämätön riski).

Tapahtuman todennäköisyys	Vähäiset seuraukset	Haitalliset seuraukset	Vakavat seuraukset
Epätodennäkönen	merkityksetön (1)	vähäinen (2)	kohtalainen(3)
Mahdollinen	vähäinen (2)	kohtalainen(3)	merkittävä (4)
Todennäköinen	kohtalainen(3)	merkittävä (4)	sietämätön (5)

Riskianalyysi esiteltiin ohjausryhmälle joulukuun kokouksessa ja ohjausryhmällä ei ollut siihen huomautettavaa – sietämättömiä riskejä ei ollut. Taulukon neljä vasemmanpuolimmaista saraketta ovat osa ohjausryhmälle esitettyä riskianalyysiä, kun taas viiden ja viimeinen sarake on loppuraporttiin kirjattu arvio kyseisen riskin hallinasta ja vaikutuksesta projektille.

4.1 Sopimukset

Sopimukset venäläisten osapuolien kanssa solmittiin joulukuun ensimmäisellä viikolla, joten sopimusten aikaansaamiseen ei liittynyt enää suoranaisia riskitekijöitä, kuten syksyllä 2012 oli pelätty. Sen sijaan sopimuksen toteuttamiseen liittyi vielä tuolloin muutamia riskitekijöitä, joita tässä yhteydessä on tuotu esille.

National Library Resourcen kanssa sovittiin tekijänoikeuksien selvittämisestä. Sopimuksessa määriteltiin osapuolien vastuut ja velvollisuudet. Samoin selvitystyön korvauksista päästiin yhteisymmärrykseen ja ne määriteltiin joulukuun 2012 alussa allekir-

joitetussa sopimuksessa. Silti tekijänoikeuksista sopimiseen liittyi riskejä, sillä tuolloin oletettiin, että mahdollisista tekijänoikeuksien korvaamisesta ja julkaisuluvan luovuttamisesta joudutaan sopimaan erikseen tapauskohtaisesti. Tekijänoikeuksiin liittyviä riskejä arvioitiin seuraavasti:

Riskin kuvaus	Projekti-päällikön arvio	Perustelu	Mahdollinen toimenpide	Toteuma ja huomautukset
Tekijän-oikeuskorva-usten loppu-summa	Vähäinen (2)	Digitoitava aineisto vanhaa ja julkaisemi-sesta pääosin yli 75 vuotta, tai se on jul-kaistu ennen 1939, joten on epäoden-näköistä että tekijän-oikeudenhaltijoiden määrä nousee suureksi.	Mikäli korvausten määrä nousee isoksi, voidaan a) budjettia allo-koida uudestaan, b) mikäli summa nousee korkeaksi, voidaan anoa täydentävää apu-rahaa, c) tai uuden arvi-oinnin perusteella jättää aineisto hankkimatta.	Ei vaikutusta projektille. Tekijänoikeuksista ei ole maksettu korvauksia, mutta Kansalliskirjasto on varannut nimellisen summan, mikäli tekijän-oikeuksienhaltijat esittä-vät vaateita.
Tekijänoikeu-den haltijat eivät myönnä lupaa julkaista aineistoa	Vähäinen (2)	Vaikka digitoitavien aineistojen joukossa paljon rinnakkais-nimekkeitä ja tarvi-taan sekä kääntäjän, että tekijän suostu-mus julkaisulupaan, niin tapausten määrä jää matalaksi.	Mikäli lupaa ei myönnetä, niin onko aineisto korvattavissa toisella materiaa-lilla? Digitoidaan, mutta aineisto jää Venäjän Kansal-liskirjaston sisäi-seen käyttöön	Ei vaikutusta projektille. Tekijänoikeudenhaltijoita ei tavoitettu National Library Resourcen selvi-tystyöstä huolimatta. Orpoteoksia koskeva lainsäädäntö mahdollisti julkaisemisen.

4.2 Venäläiset organisaatiot

Venäläisten yhteistyökumppaneiden toimintaan liittyi joitain riskejä, jotka olivat olleet tiedossa jo projektin alusta lähtien, mutta myös sellaisia, joita ilmeni pilotoinnin aikana.

Riskin kuvaus	**Projektipäällikön arvio**	**Perustelu**	**Mahdollinen toimenpide**	**Toteuma ja huomautukset**
Digitointia ei saada valmiiksi aikataulussaan / lainkaan	kohtalainen (3)	Venäjän Kansalliskirjastossa on käynnissä parhaillaan organisaationmuutos – vähentämistarve on noin 350 henkeä ja tämä vähennys saattaa vaikuttaa aikatauluihimme. Onko Pietari lupautunut toimittamaan aineiston aikaisemmin kuin pystyy toteuttamaan?	Vahvistettua tietoa vähennysten vaikutuksesta hankkeelle ei ole, mutta keskusteluista ja toimenpiteistä (mm. skannerin hankkiminen) päätellen he ovat sitoutuneita projektiin. Vaikutusta lähinnä sanomalehtien skannaamiseen? Yhteistyön koordinointi kärsii.	Riski toteutui osittain ja vaikutus projektille oli kohtalainen. Varsinaiset digitoinnit oli tehty ajoissa, mutta jälkikäsittely Pietarissa viivästytti aineiston toimittamista yli sovitun aikarajan. Riskiä ei voitu ennakoida, sillä Venäjän kansalliskirjastosta vakuuteltiin koko prosessin ajan, että kaikki aineisto tullaan toimittamaan ajoissa, mutta näin tapahtui ainoastaan monografioiden osalta. Kun palvelu julkistettiin 6. kesäkuuta, vain noin kymmenesosa sanomalehtiaineistosta oli Fenno-Ugricassa. Viimeiset sanomalehtiaineistot kuvailtiin tietokantaan syyskuussa, yli neljä kuukautta sovitun toimitusajankohdan jälkeen. Syynä tähän viivästykseen oli sen, että sanomalehtiaineiston OCR-työ kesti oletettua pidempään ja syynä tähän oli työvaiheen aliresursointi, osaamiskapeikko ja työn huono koordinointi Pietarissa.
Aineiston laatu digitointiin kelpaamatonta	vähäinen (2)	Kaikkea sanomalehtiaineistoa ei pystytä digitoimaan sen heikon kunnon vuoksi. Noin 2000 sivua jää toistaiseksi digitoimatta.	Aineiston entisöinti maksaisi 513 000 RUR + tuotantokulut, eli arviolta noin 943 000 RUR / 23500 EUR. Tämänhetkiseen budjettiin kuluerä oli liian suuri, mutta voisiko tämän aineiston digitointiin hakea täydentävää rahoitusta? Olisi kartoitettava saadaanko materiaali muualta joko alkuperäisestä skannaamalla tai mikrofilmiltä digitoimalla? Vaatisi joka tapauksessa lisärahoitusta.	Vaikutus projektille vähäinen. Valtaosa aineistoista pystyttiin digitoimaan ja vain osa aineistosta vaati konservointia. Viimeiset konservointia vaatineet aineistot toimitetaan Kansalliskirjastolle huhtikuun 2014 loppuun mennessä. Osa digitoitavaksi esitetystä aineistoista tosin oli kärsinyt vuosia aikaisemmin vesivahingoista, eikä niitä sen vuoksi voitu digitoida. Tämä ei ollut tiedossa aineistoa valitessa.
National Library Resourcen asema vahvistuu talvella 2013 tekijänoikeusjärjestönä	vähäinen (2)	National Library Resource tulee jatkossa lisensoimaan / myöntämään tekijänoikeuden haltijoiden puolesta luvan nyk. Venäjän federaation alueella julkaistuun aineistoon ja vie kansalliskirjastoilta mahdollisuuden sopia tekijänoikeudenhaltijoiden kanssa suoraan.	Keskusjohtoisuus lisääntyy, mutta meillä on nyt jo kokemusta NatLibin kanssa toimimisesta. Huonoa se, että joudumme käymään Moskovan kanssa keskusteluja, vaikka aineisto digitoitaisiin muualla. Pitäisi saada aikaan pieni projekti esim. Udmurtian kanssa ennen kuin laki astuu voimaan. Keskeisin vaikutus kuitenkin vuosille 2013–2016.	Vaikutus projektille vähäinen. National Library Resourcen asemaa venäläisen kirjastoalan tekijänoikeusjärjestönä asemaa ei oltu vahvistettu duumassa projektin loppuun mennessä. Venäläisten partnerien mukaan asia on edelleen käsittelyssä.

Tekijänoikeuksienhaltijoiden myöntämän luvan jakaminen ei onnistu	vähäinen (2)	On allekirjoitettu aiesopimus, minkä perusteella Kansalliskirjasto saa luovuttaa edelleen kolmansille osapuolille oikeuden julkaista aineistoa, koskee ml. Venäjän Kansalliskirjastoa. Kyseessä on kuitenkin vain aiesopimus, mikä ei ole sitova millekään osapuolelle.	Mahdollinen konflikti Venäjän Kansalliskirjaston kanssa vältettävissä – olemme sitoutuneet avoimeen yhteistyöhön ja noudattamaan sekä sopimuksia että lainsäädäntöä. Pietarilla ei vastustusta, mutta pitää huomioida Pietarin heikko asema lisensoinnin suhteen. Vaikuttaa myös vuosina 2013–2016, mikäli digitoidaan Pietarista.	Vaikutus projektille vähäinen. Digitoidun aineiston jakaminen kolmansille osapuolille onnistuu, sillä aineisto on tekijänoikeuksienselvityksen myötä public domain Venäjän lainsäädännön mukaan. Digitointeja koskeva tekijänoikeuksien selvitys tehdään jatkossakin yhteistyössä NatLibin kanssa.

4.3 Budjetti

Budjettiin kohdistuvat riskitekijät kohdistuvat lähinnä kuluerien uudelleen kohdistamiseen ja mahdollisten tekijänoikeuskorvausten määrään.

Riskin kuvaus	Projektipäällikön arvio	Perustelu	Mahdollinen toimenpide	Toteuma ja huomautukset
Tekijänoikeuskorvausten suuruus ylittää budjetin	kohtalainen (3)	Emme tiedä mikä tulee olemaan tekijänoikeudenhaltijoiden korvausvaatimusten taso, mikäli sellaisia ylipäänsä ilmenee. Selvitystyön aikana ei saatu tietoja selvitystyön suuruudesta, eikä mahdollisista korvausmääristä, vaan tämä kuluerä jouduttiin arvioimaan ilman parempaa tietoa. Kustannusten odottamaton kasvaminen euroalueen kriisin tai valuuttakurssien heittelyn seurauksena	National Library Resource on ottanut vastuulleen korvata selvitystyön jälkeen esitettävät tekijänoikeuskorvausvaatimukset, jolloin Kansalliskirjasto ei joudu korvaamaan mitään jälkikäteen. Digitointi ja selvitystyö laskettu siten, että noin 5% kulueristä vielä käyttämättä. Budjetissa muuallakin liikkumavaraa (mm. koulutuksesta), jota voidaan allokoida toiseen toimintoon, mikäli kustannukset kasvavat. Mikäli tekijänoikeuksien korvaamisesta koituvat kustannukset kasva-	Ei vaikutusta projektille. Tekijänoikeuksienhaltijoille ei ole maksettu korvauksia, sillä heitä ei ole selvityksen yhteydessä paikannettu. ks. myös luku Tekijänoikeudet ja niiden selvittäminen.

			vat yli budjetin, neuvoteltava rahoittajan kanssa mahdollisesta lisärahoituksesta ja mikäli lisärahoitusta ei saada, tulee keskustella tutkijoiden kanssa siitä, mitä aineistoja otetaan mukaan.	
Henkilöstö/budjetoitu työnmäärä ei riitä hankkeen toteuttamiseen	vähäinen (2) / merkittävä (4)	Selvitysvaiheessa arvioitu työnmäärä ylittyy ja projektille tehtävää työtä ei voida tehdä virkatyönä. Tietojärjestelmäasiantuntijan palkkaaminen ja yhteistyö Kirjallisuuspankin kanssa parantaa nykytilannetta. Projektipäällikölle allokoitu työmäärä riittämätön hankkeen täysimittaiselle toteuttamiselle.	Rekrytointikiellon vuoksi jouduttiin tekemään enemmän virkatyötä kuin oli ajateltu – riittävätkö korvaavalle työlle allokoidut varat pilottikauden loppuun? Riskeinä resurssien ja osaamisen poistuminen hankkeesta. Jos resurssit lähtevät, mistä korvaavat resurssit tilalle? Riskinä hankkeen kaatuminen. Projektipäällikkö tehnyt suunniteltua enemmän töitä sopimusvaiheessa ja jäljellä olevilla resursseilla voidaan keskittyä lähinnä pilotin läpiviemiseen, mutta kehittämiseen ei suuriakaan mahdollisuuksia. Selvitystyön jälkeen (maalis-huhtikuussa) uusi arvio työajan kohdentamisesta ja mahdol-	Vaikutus projektille kohtalainen. Helsingin yliopiston rekrytointikielto hidasti projektiin rekrytoidun tietojärjestelmäasiantuntijan palkkaamista yli kolmella kuukaudella. Rekrytointilupa saatiin marraskuussa 2012 ja haastattelut suoritettiin saman kuun lopussa. Tehtävään valittu henkilö aloitti työssään joulukuun puolessa välissä. Tämän viivästyksen vaikutukset ovat näkyneet ennen kaikkea OCR-editorin kehittämisessä. Vaikka Kansalliskirjaston kirjasto verkkopalveluissa aloitettiin editorin kehittäminen virkatyönä, niin editorin täysipainoinen suunnittelu voitiin aloittaa vasta tietojärjestelmäasiantuntijan aloitettua työnsä. Projektipäällikön käyttämä työaika ylitti suun-

			lisista uusista allo-koinneista.	nitteluvaiheessa lasketun työaika-arvion lähes kaksinkertaisesti. Pilot-tihanke olisi voinut epä-onnistua täydellisesti ellei Kansalliskirjasto olisi kokopäiväistänyt projektipäällikköä huhti-kuun 2013 alusta alkaen.

4.4 Aikataulut

Hankkeen toteuttamisen kannalta tietyt tehtävät ja niihin liittyvät aikataulut ovat kriittisiä. Osa aikataulutukseen liittyvistä riskeistä syntyy venäläisten toimijoiden asettamista rajoista, kuten selvitystyön kestosta, kun taas Kansalliskirjaston sisällä tehtävä työ ja sen aikatauluttaminen voi muodostaa riskin tutkijayhteistyön täysimittaiselle toteuttamiselle.

Riskin kuvaus	**Projektipäällikön arvio**	**Perustelu**	**Mahdollinen toimenpide**	**Toteuma ja huomautukset**
OCR-editoria ei saada valmiiksi ajoissa	kohtalainen (3)	OCR-editoria ei saada käyttöön kun aineisto voidaan asettaa julkisesti saataville. Rekyrtointikiellon vuoksi työ pääsee alkamaan yli kolme kuukautta myöhemmin kuin oli suunniteltu.	OCR-editoria on kehitetty virkatyönä ja lähtökohdat sen toteuttamiseen talven ja kevään aikana ovat olemassa. Se mitkä editorin toiminnot ovat tutkijoiden käytössä ja mistä alkaen, on vielä epäselvää. Tahtotila editorin kehittämiseen on olemassa. Jos OCR-editoria ei saada aikaiseksi, niin vaihtoehtona on mm. sen tuottamisen ulkoistaminen.	Vaikutus projektille on ollut ennakoitua suurempi. OCR-editoria ei ole vielä (lokakuu 2013) saatu tutkijoiden käyttöön kokonaisuudessaan. Yksi työtä viivästyttäneistä syistä on ollut Helsingin yliopiston rekrytointikielto, minkä vuoksi editorin kehittämisen aloitus viivästyi. Toinen keskeinen tekijä on ollut suunnittelu- ja koodaustyön aliresursointi - siitäkin huolimatta, että editorin kehittämiseen on laitettu myös virkatyötä. Yhden tietojärjestelmäasiantuntijan sijaan olisi tarvittu enemmän tai jaettuja resursseja, jolloin työvaiheita olisi voitu jakaa useamman tekijän kesken.
Selvitystyön kesto viivästyttää hanketta	merkittävä(4)	Selvitystyö tulee saattaa päätökseen 90 työpäivän kuluessa, sopimuksen allekir-	90 työpäivää on pitkä aika ja jos koko aika käytetään, mikä on erittäin todennäköis-	Vaikutus projektille oli arvioitua vähäisempi. Tekijänoikeuksienselvitys

		joittamisesta (4.12.2012) lähtien.	tä, työ saadaan päätökseen vasta maalis- ja huhtikuun vaihteessa. Tekijänoikeuksista vapaa materiaali voitaneen julkaista heti selvitystyön päätyttyä, mutta jos oikeudenhaltijoita ilmenee, pitää heidän kanssaan vielä sopia tekijänoikeuksien luovuttamisesta, korvauksista jne. Mitä työtä tutkijat voivat sillä aikaa tehdä?	valmistui sovitussa ajassa ja Kansalliskirjasto hyväksyi National Library Resourcen tekemän selvityksen huhtikuun alussa. Selvitystyö ja sen kesto eivät hidastaneet projektia ja aineistojen saattamista avoimeen ja tutkijoiden käyttöön.
Aineiston digitointi ja käyttöönsaattaminen viivästyy	merkittävä (4)	Pietarissa olevat resurssit ovat vajaat ja he ovat suostuneet kovin tiukkaan aikatauluun. Pystyvätkö pitämään digitoinnin tason korkealla? Sanomalehtiaineiston käsittely ja konservointi vie aikaa. Ei riittävästi laitteistoa?	Pietari ei tule pysymään aikataulussa, mutta viivästysten vaikutuksen hankkeen toteuttamiselle todennäköisiä, mutta kuitenkin pieniä. Viivästykset tulevat kohdistumaan sanomalehtiaineistoon, jota ei pystytä digitoimaan yhtä paljon kuin Pietari on itse arvioinut, sillä työskentelevät kahdella sanomalehtien digitointiin soveltuvalla skannerilla, mutta toisen skannerin (A1-kokoinen) tuottama aineisto joudutaan yhdistämään toisiinsa, mikä vie aikaa ja mahdollisesti heikentää laatua. Testiaineisto tehty toisinaan viimeisen päälle, toisinaan huonosti. Mihin digitoinnin taso asettuu ja kuinka paljon voimme hyväksyä aineistoja?	Riski toteutui osittain ja vaikutus projektille oli kohtalainen. Varsinaiset digitoinnit oli tehty ajoissa, mutta jälkikäsittely Pietarissa viivästytti aineiston toimittamista yli sovitun aikarajan. Riskiä ei voitu ennakoida, sillä Venäjän kansalliskirjastosta vakuuteltiin koko prosessin ajan, että kaikki aineisto tullaan toimittamaan ajoissa, mutta näin tapahtui ainoastaan monografioiden osalta. Kun palvelu julkistettiin 6. kesäkuuta, vain noin kymmenesosa sanomalehtiaineistosta oli Fenno-Ugricassa. Viimeiset sanomalehtiaineistot kuvailtiin tietokantaan syyskuussa, yli neljä kuukautta sovitun toimitusajankohdan jälkeen. Syynä tähän viivästykseen oli sen, että sanomalehtiaineiston OCR-työ kesti oletettua pidempään ja syynä tähän oli työvaiheen aliresursointi ja huono työn koordinointi Pietarissa.

5 Julkaisujärjestelmä

Fenno-Ugrica (fennougrica.kansalliskirjasto.fi) on toteutettu DSpace-ohjelmistolla. DSpace on avoimen lähdekoodin sovellu digitaalisten aineistojen hallinnointiin. Kansalliskirjasto käyttää sitä eräiden omien aineistojensa hallintaan sekä tarjoaa sen avulla tuotettuja maksullisia julkaisuarkistopalveluja.

Pilottihankkeessa räätälöidyn DSpace-instanssin avulla edistetään digitoidun aineiston saavutettavuutta ja käyttöä myös ulkomailla, niin kansainvälisen tiedeyhteisön kuin Venäjällä asuvien suomensukuisten kielten puhujien keskuudessa. Fenno-Ugrica toteutettiin myös venäjänkielisenä ja sen käyttöliittymä voidaan mahdollisesti julkaista myöhemmin myös muilla kielillä. Viestinnän ja tunnettuuden edistämiseksi pilottihankkeen DSpace-instanssille luotiin räätälöity ja tunnistettava ulkoasu.

-

Kuva 1. *Fenno-Ugrican venäjänkielinen käyttöliittymä*

5.1 Fenno-Ugrica –kokoelma julkaisujärjestelmänä

Fenno-Ugrica on toteutettu virtuaalipalvelimeen, jota ajetaan Kansalliskirjaston Vmware-ympäristössä. Virtuaalipalvelimen levytilan koko on 337 Gt, ja siitä on tällä hetkellä käytössä noin 70 %. Fenno-Ugrican käyttöliittymä ja käyttöohjeet on toteutettu suomeksi, englanniksi ja venäjäksi. Sille on suunniteltu oma graafinen ulkoasu. Erityispiirteenä Fenno-Ugrican on latinalaisten ja kyrillisten julkaisujen rinnakkaisuus. Nimekkeiden ja tekijöiden selauksessa on käytössä rinnakkaiset listaukset molemmille järjestelmille. Lisäksi käyttöliittymään on toteutettu mahdollisuus selata aineistoja kielen mukaan.

Käyttöön asettamista varten aineistot on täytynyt kuvailla. Venäjän kansalliskirjastolla on saa-dusta aineistosta kuvailutietoja, mutta niiden automaattinen hyväksikäyttäminen osoittautui vaikeammaksi kuin uudelleen kuvailu. Lisäksi aineistojen syöttöprosessiin, jossa mm. lisätään aineistoon metatietoja on tehty aineiston edellyttämiä räätälöintejä. Itse julkaisut asiakas saa käyttöön PDF-tiedostoina. Käyttöliittymään upotettua lukulaitetta ei ainakaan tässä vaiheessa ole katsottu tarpeelliseksi toteuttaa, koska se ei tuottaisi asiakkaalle merkittäviä uusia ominaisuuksia tai parempaa lukukokemusta. Kaikki saatu aineisto on siis tällä hetkellä asetettu yleisökäyttöön siinä kunnossa kuin ne on saatu; kun tunnistetun tekstin korjaaminen on valmis, kukin dokumentti korvataan parannetulla versiolla.

Tällä hetkellä Fenno-Ugricasta ladataan tiedostoja keskimäärin 900 kertaa kuukaudessa. Kansalliskirjaston palvelut näkyvät erittäin hyvin hakukoneissa; aineiston luonteen huomioon ottaen käyttö on vähintäänkin runsasta.

5.2 Fenno-Ugrican metadata

Pilottihankkeessa digitoitujen aineistojen kuvaillussa on käytetty DublinCore kuvailuformaat-tia. Venäjän Kansalliskirjasto käyttää aineiston kuvailuformaattina RusMarcia ja Suomessa painetun aineiston kuvailussa on käytössä Marc21-formaatti. Ennen kuin projektissa digitoitu aineisto voitiin kuvailla Fenno-Ugricaan, tuli päättää, mitä DC-kenttiä tarvitaan tässä projektissa ja mitä tietoa mihinkään kenttään tallennetaan. DC-kentät määriteltiin yhdessä Marina Ivanovan, Jussi-Pekka Hakkaraisen ja Ulla Ikäheimon kesken.

Projektin alkuvaiheessa kirjastonhoitaja Marina Ivanova tutustui tarkemmin venäläiseen RusMarc-formaattiin ja hän osallistui DSpace-koulutukseen, joka järjestettiin 3.10.2012 Kansalliskirjaston Kirjastoverkkopalvelussa. Pystyäkseen työskentelemään DSpacessa käytettävällä metadataformaatilla, Ivanova perehtyi myös suomalaisiin ja ulkomaalaisiin DublinCore-formaatin käyttöoppaisiin ja käytöntöihin.

DublinCore -formaatti on käytetty digitaalisina syntyneiden verkkodokumenttien kuvailuun jo 1990-luvusta alkaen, joskin digitoitujen aineistojen kuvailussa se ei ole vielä vakiintunut ja DC-kenttiä käytetään vaihtelevasti. Esimerkkinä voi mainita julkaisija-kentän. DC-ohjeen mukaan julkaisija-kentän tarkoituksena on identifioida se taho joka on asettanut kyseisen tallenteeseen käyttöön. Monissa DSpace-järjestelmissä kuvailtaessa digitoituja dokumentteja tähän kenttään on kuitenkin tallennettu alkuperäisen (painetun) dokumentin julkaisutietoja.

Kuvailtu aineisto sisältää 156 digitoitua monografiaa inkeroisen, vepsän, ersän, mokšan, šokšan, niittymarin ja vuorimarin kielillä. Inkeroisen ja vepsän kirjaimisto on latinalainen, mui-den kielten kirjaimisto on kyrillinen. Kyrillinen aineisto on kuvailtu käyttäen sekä alkuperäisiä merkkejä, että translitterointia. Translitterointistandardina on valittu ISO 9 -standardi (Transli-teration of Cyrillic characters into Latin characters – Slavic and non-Slavic languages), koska se on käytetyin suomalaisissa kirjastoissa. Translitteroinnissa runsaasti käytetyt erikoismerkit eli diakriitit vaativat tarkkuutta ja kärsivällisyyttä, koska normaalista näppäimistöstä niitä ei saa ja ne piti lisätä kopioimalla eri lähteistä.

Kirjastoalalla on paljon käytetty MARC-kielikoodeja, jotka perustuvat Library of Congressin ylläpitämään ISO-standardiin ISO 639-2. Tässä projektissa niitä ei voinut kuitenkaan hyödyn-tää, koska kaikille suomalais-ugrilaisille kielelle siinä ei ole omaa kielikoodia. Erilaisista kan-sainvälisistä kielikoodistandardista valitsimme ISO 639-3 standardin (Sukukielten digitointiprojekti 070714.docx), koska se on täydellisempi ja sisältää kielikoodit kielille, joitta tarvitsemme tässä projektissa.

Aineiston monikielisyys asetti haasteita kuvailuprosessille. Esimerkiksi tekijän nimimuodon valinnassa piti päättää, valitaanko alkuperäinen venäjänkielinen nimimuoto - Тетюрев vai käännöksissä esiintyvät muodot – Тьэтьурэв, Тэтьурэв, Tetjurev ja Tetyrev? Kirjastotietokannoissa nämä ongelmat voidaan ratkaista luomalla auktorisoituja nimimuotoja, joiden avulla on mahdollista saada saman tekijän kaikki teokset yhdellä tekijähaulla. DSpace -ympäristö ja DublinCore-formaatti eivät anna tällaiseen auktorisoitujen nimimuotojen käyttöön mahdollisuutta. Tässä projektissa teoksien määrä on kuitenkin vähäinen ja haettu nimimuoto löytyy selaamalla tekijälistaa, joka ei ole pitkä. Projektissa digitoitu aineisto sisältää oppikirjoja ja sanakirjoja. Asiasanoituksen suhteen yleisiä ohjeita ja sovittuja käytänteitä oli helpompi seurata, sillä termeinä käytettiin vain ko. kieltä ja oppiainetta. Sisällönkuvailu tehtiin suomen- ja venäjänkielisillä asiasanoilla.

6 OCR-editori kielentutkimuksen tueksi

Sukukielten digitointiprojektilla on kiinnekohtia myös kieliteknologiseen tutkimukseen, sillä projektin yhdeksi päämääräksi voidaan laajasti ajatella digitaalisten kirjasto- ja arkistoaineistojen käyttötapojen ja käytettävyyden parantaminen. Projektissa on edistetty suomalais-ugrilaisen aineiston käyttöön saattamisen lisäksi menetelmiä, joilla digitoitua raakadataa voidaan jalostaa entistä käyttökelpoisemmiksi aineistoiksi ja joilla aineistoa voidaan hyödyntää.

Sukukielten digitointiprojektissa näillä menetelmillä tarkoitetaan digitoidun aineiston OCR-tunnistuksen lisäämistä, tunnistetun tekstimassan palstoittamista sekä ennen kaikkea kielenkorjaukseen tarkoitetun OCR-editorin kehittämistä, jonka avulla voidaan digitoinnin ja OCR-tunnistuksen yhteydessä jääneitä virheitä korjata tehokkaasti ja talkoistamalla. Kansalliskirjaston pilottivaiheen aikana kehittämä OCR-editori ainoa laatuaan maailmassa ja se tukee myös digitointialan kokonaisvaltaista kehittymistä.

6.1 Yleiskatsaus tekniseen ratkaisuun

Projektin teknisessä toteutuksessa on kolme pääkomponenttia:

1) aineiston vastaanotto sisällön toimittajalta
2) järjestelmä, jonka avulla aineisto esitetään yleisölle
3) järjestelmä, jonka avulla aineiston tekstitunnistusta korjaillaan ja jossa siihen liittyviä prosesseja hallinnoidaan.

6.2 Aineiston vastaanotto

Projektin aineisto on digitoitu Pietarissa ja se siirrettiin Kansalliskirjastoon Pietarissa olevalta FTP-palvelimelta. Aineisto on tuotettu palvelimelle sovitunrakenteisina kokonaisuuksina, jotka ovat zip-pakattu. Yksi paketti sisältää yhden monografian tai yhden lehden numeron. Kukin paketti sisältää PDF-tiedoston, johon on upotettu automaattisen tekstintunnistuksen tulos, sekä kunkin sivun erillisenä kuvatiedostona. Kansalliskirjasto on muodostanut PDF-tiedostoista ALTO XML -tiedoston. ALTO on vanhojen aineistojen digitoinnissa yleisesti käytetty tapa ilmaista sisältö rakenteisessa XML-muodossa. Tiedostosta käy ilmi dokumentin rakenne, mm. kaikkien sanojen tarkat sijaintikoodinaatit. Tässä vaiheessa myös ALTO XML-tiedosto siis sisältää tunnistetun tekstin tarkistamattomana. Paketissa olevat PDF-tiedostot esittävät julkaisun facsimilena, jossa on taustalla upotettuna automaattisen tekstin-

tunnistuksen tulosteksti; nämä tiedostot on kopioidaan Fenno-Ugrica –julkaisu-.arkistoon julkista esittämistä varten. Aineisto varmistetaan nauhoille. Tähän mennessä aineiston koko on kaikkiaan noin 1,8 teratavua. Jäljellä olevien latausten toteuduttua tilan tarve on noin 4 teratavua.

6.3 OCR-editori

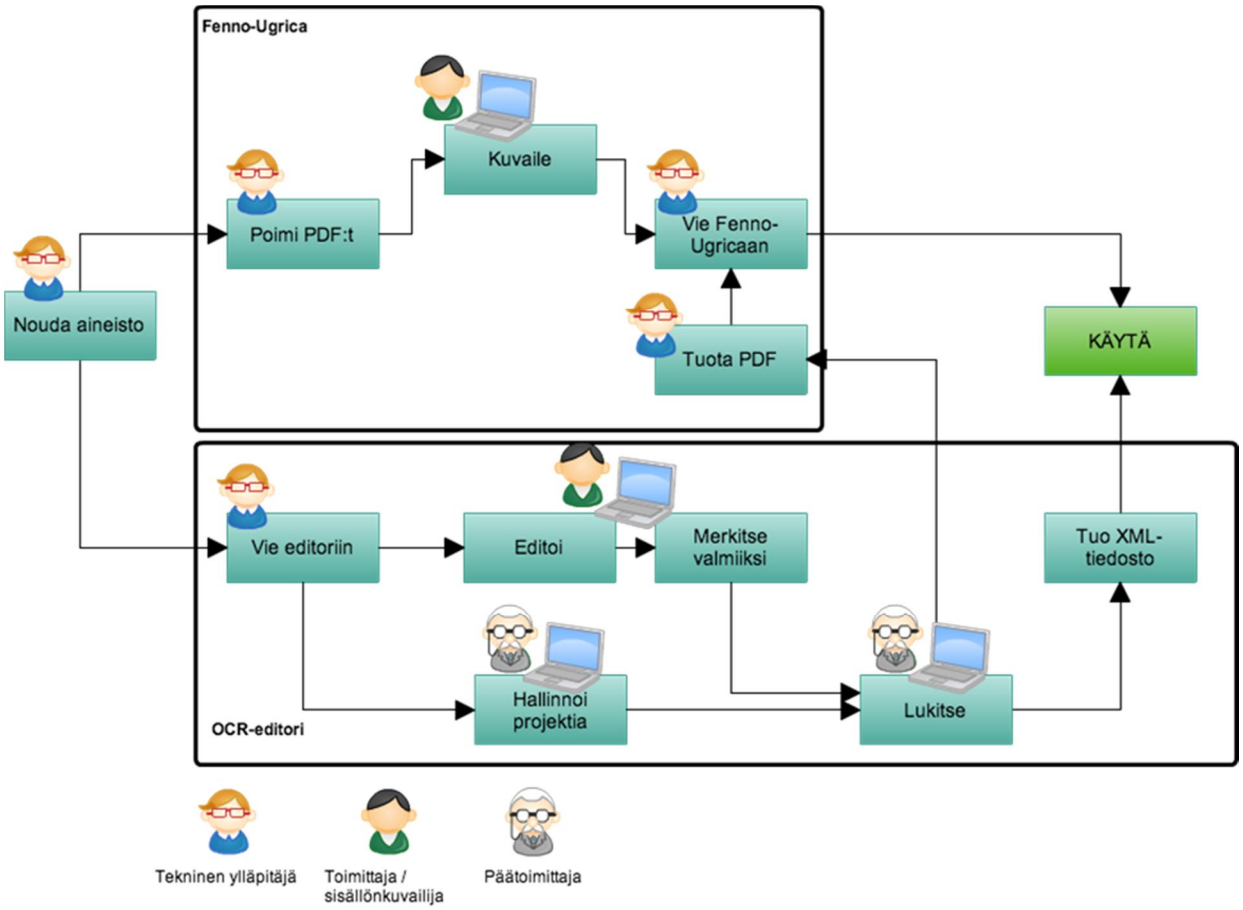
OCR-editori koostuu kahdesta pääosasta:

1) editorin käyttöliittymästä, jota tekstin korjaajat käyttävät sekä
2) sen taustalla olevasta järjestelmästä, jossa hallinnoidaan tietokantoja, aineistoja ja niiden versiointia, käyttäjiä, selausnäkymiä ja muita editoinnin vaatimia toimintoja.

Käyttöliittymä on toteutettu JavaScriptillä ja taustajärjestelmä Pythonilla. Tiedonsiirto tapahtuu JSON API:lla toteutetun REST-rajapinnan avulla. Kansalliskirjasto tulee piakkoin julkaisemaan OCR-editorin avoimena lähdekoodina. OCR-editoriin ladataan saaduista paketeista ALTO XML -tiedosto ja kuvatiedostot sekä paketeista saatava metadata; nimeke, tekijä ja pääasiallinen kieli. Näistä muodostetaan aineistoluettelo, josta aineisto voidaan valita avattavaksi editointikäyttöliittymässä.

6.4 OCR-editorilla työskentely

- Toimittaja valitsee haluamansa dokumentin aineistoluettelosta.
- OCR-editorissa toimittaja näkee näytöllä rinnakkain dokumentin yhden sivun facsimilena ja sen konetunnistetun tekstin.
- Toimittaja voi pitää kuvan ja tekstin rinnakkain tai päällekkäin
- Kuvaa voi liikutella hiirellä, pienentää ja suurentaa
- Klikkaamalla sanaa joko kuvassa tai tekstissä, sana korostuu molemmissa versioissa
- Tekstiä voi korjailla tavanomaiseen tapaan. Korjatut sanat saa näkymään korostettuina.
- Erikoisten merkistötarpeiden takia toimittaja saa käyttöönsä myös merkkivalikon, josta valitsemalla tekstiin voi lisätä merkkejä, jotka ovat hankalia käytössä olevalla näppäimistöllä.
- Toimittaja voi valita yksittäisen sanan tai tekstinosan ja merkitä sen kielen.
- Toimittaja voi tallentaa tekemänsä muutokset
- Toimittaja voi palata aiemmin tallennettuun versioon (työn alla per 17.10.2013).
- Toimittaja voi merkitä dokumentin mielestään valmiiksi korjatuksi (työn alla per 17.10.2013.



Kuva 2. *Työvuo Fenno-Ugricassa ja OCR-editorissa*

6.5 OCR-editorin hallinnointi

Hallinnointitehtävillä tarkoitetaan editointiprojektin päätoimittajan tehtäviä. Niitä ovat mm.

- Aineiston kokoelmarakenteen luominen
- Uusien käyttäjien luominen
- Käyttöoikeuksien antaminen käyttäjille
- Aineistojen lukitseminen enemmiltä muutoksilta.

Taustajärjestelmä on suunniteltu niin, että se mahdollistaa toisistaan täysin irralliset projektit, so. aineistot ja käyttäjät eivät näy yli projektin rajojen. Käyttöoikeuksien hallinnointi perustuu kokoelmarakenteeseen. Toimittaja voi editoida aineistoja ainoastaan niissä kokoelmissa, joihin hänelle annetaan riittävät oikeudet. Päätoimittaja voi myös delegoida hallinnointioikeuksia. Hallintokäyttöliittymän toteutus on vielä työn alla (per 17.10.2013).

6.6 Aineistojen tuominen OCR-editorista

Aineistoja voi tuoda OCR-editorista sivu tai dokumentti kerrallaan. Käytännössä jokaista sivua vastaa yksi ALTO XML -tiedosto; tarvittaessa ne voidaan esittää peräkkäin. Editorin suunnittelun yhteydessä on keskusteltu siitä, missä muodossa aineiston pitäisi olla saatavissa. On ilmeistä, että XML-tiedosto on jatkokäsittelyn kannalta kiitollisin. Sen käyttö edellyttää hieman taitoa mutta on kaiken kaikkiaan vain tekstinkäsittelyä. XML-tiedostosta voi varsin helposti tuottaa ihmisluettavan version, siis pelkän tekstin; tai sanaluettelon; tai muutakin.

7 Viestintä

Pilottiprojektin viestintä on ollut projektipäällikön vastuualueella ja hän on vastannut viestinnällisistä sisällöistä ja tietojen paikkaansa pitävyydestä. Hänen käytössään ovat olleet Kansalliskirjaston ja Helsingin yliopiston yleiset viestintäkanavat. Viestintää on harjoitettu pääsääntöisesti suomeksi, mutta tarvittaessa sisältöjä on myös käännetty englanniksi ja venäjäksi.

7.1 Verkkosivut

Kansalliskirjaston verkkosivujen nk. siniselle kaistalle, jonne on kerätty tietoja Kansalliskirjaston kokoelmista ja palveluista, luotiin Sukukielten digitointiprojektille oma osio Digitoitujen kokoelmien -välilehden alle. Verkkosivut löytyvät osoitteesta: [http://www.kansalliskirjasto.fi/kokoelmatjapalvelut/digitaalisetkokoelmat/finnougric.html]

Verkkosivuilla tarjottiin perustietoja hankkeesta ja sieltä pystyy lukemaan mm. pilottivaiheen projektisuunnitelman sekä digitoitavien aineistojen listat. Sisällöt toteutettiin myös englannin- ja venäjänkielisinä.

7.2 Kansalliskirjasto-lehti

Sukukielten digitointiprojektin pilottihanketta esiteltiin projektipäällikön kirjoittamien juttujen avulla myös Kansalliskirjasto-lehden numeroissa 4/2012 ja 4/2013 .

7.3 Blogi

Sukukielten digitointiprojektin blogi on ollut tärkeä viestinnän väline, sillä se on tarjonnut Kansalliskirjaston verkkosivuja joustavamman foorumin keskustella ja tiedottaa hankkeen etenemisestä. Blogitekstejä on kirjoittanut pääsääntöisesti projektipäällikkö, mutta myös tutkijat ja venäläiset yhteistyökumppanit ovat voineet jakaa näkemyksiään blogissa. Blogi on englanninkielinen ja se sijaitsee osoitteessa https://blogs.helsinki.fi/fennougrica/

Kuva 3. *Sukukielten digitointiprojektin blogi*

7.4 Esitelmät

Projektipäälliköllä on ollut mahdollisuus esitellä hanketta erilaisissa kirjastoalan, tieteenhistorian ja digitaalisen kulttuurin tapahtumissa ja seminaareissa. Esitelmiä on pidetty mm. seuraavis-sa tilaisuuksissa:

- 8.8.2012 IFLA, Helsinki
- 22.9.2012 Venäjän kansalliskirjasto, Pietari
- 13.3.2013 Kirjaston tori, Helsinki
- 10.4.2013 Nekrasov-kirjasto, Moskova
- 12.4.2013 Udmurtian tasavallan kansalliskirjasto, Iževsk
- 16.5.2013 Karjalan tasavallan kansalliskirjasto, Petroskoi
- 5.6.2013 Variantti-kollokvio 2013: KORPUS-TIETOKANTA–EDITIO, Helsinki
- 6.6.2013 Fenno-Ugrica –kokoelman julkaiseminen, Helsinki
- 26.7.2013 International Congress of History of Science, Technology and Medicine, Manchester
- 23.10.2013 Viron kielen instituutti, Tallinna
- 14.11.2013 Historiallisen yhdistyksen seminaari, Helsinki.

7.5 Sosiaalinen media

Kansalliskirjaston Facebook-sivu toimi hyvänä viestintäkanavana Sukukielten digitointiprojektille. Pienemmälle ja rajatummalle yleisölle hankkeen etenemisestä tiedotettiin myös projektipäällikön oman Twitter-tilin kautta, jossa hankeviestintä tavoitti ainakin jonkin verran asiantuntijayleisöä nk. digital humanities -sektorilta. Hankkeen etenemisestä ilmoitettiin niin Facebookissa kuin Twitterissä uutisluontoisesti, eli ilmoitettiin uusien aineistojen saatavuudesta ja linkitettiin hankkeen blogiin. Uutisia hankkeen edistymisestä jaettiin sidosryhmien (mm. Uralistica, M.A. Castrénin seura, Sukukansojen ystävät jne.) Facebook-sivuilla.

7.6 Kick off –tilaisuus

Fenno-Ugrica -kokoelma julkaistiin 6.6.2013 Kansalliskirjaston Auditoriossa. Tilaisuudessa pi-dettiin kolme esittelyä: projektipäällikkö Jussi-Pekka Hakkarainen kertoi yhteistyöstä yleistietoja hankkeesta ja esitteli kokoelmaa, tutkija Jack Rueter puhui aineistoon kohdistuvasta tutkimuksesta ja tietojärjestelmäasiantuntija Juho Vuori esitteli OCR-editoria. Tilaisuuteen osallistui paikanpäällä yhteensä 67 henkilöä ja etäyhteyden välityksellä tilaisuutta seurasi yhteensä 43 henkilöä, niin Suomesta kuin Venäjältäkin.

Fenno-Ugrica -kokoelman julkaisemisesta lähetettiin STT:n kautta tiedotusvälineille suomen-, englannin- ja venäjänkielinen lehdistötiedote, mutta yksikään suomalainen media ei huomioinut kokoelman julkaisua uutisissaan. Yhtenä syynä medianäkymättömyydelle oli, että tuolla samaisella viikolla ja samana päivänä julkaistiin useita avoimen datan palveluita ja julkisuus täyttyi pikemminkin Ilmatieteenlaitoksen ja Yleisradion hankkeista kuin Fenno-Ugrica -kokoelmasta: http://essetter.blogspot.fi/2013/06/det-regnar-data.html#.UndrP1N9H-5

8 Budjetti

Pilottihankkeen kustannukset vuosille 2012–2013 muodostivat Koneen Säätiöltä haettavan apurahan suuruuden. Koneen Säätiön myöntämän apurahan suuruus oli 356 040 euroa. Pilotti-hankkeen kustannukset esiteltiin maaliskuussa 2012 Koneen Säätiölle jätetyssä apurahahake-muksessa laskemina, jotka perustuivat selvitystyön aikana tehtyihin arvioihin.

Vaikka arviot lähtökohtaisesti pitivät paikkansa hankkeen aikana, niin kaikkia kustannuseriä, kuten tekijänoikeuksista koituvia kustannuksia, ei tuolloin selvitystyön yhteydessä voitu esittää yksityiskohtaisesti, sillä selvitysvaiheessa ei ollut tietoa mahdollisista tekijänoikeuskorvauksien määrästä ja selvitystyön kustannuksista. Todellisista kustannuksista saatiin parempi ja yksityiskohtaisempi käsitys pilotti-hankkeen aikana ja budjetointiin liittyvät muutokset tuotiin aina ohjausryhmän keskusteltavaksi.

Projektilla ei ollut Kansalliskirjaston omarahoitusosuutta, sillä hankkeessa ei tuotettu aineistoja kirjaston omista kokoelmista. Omarahoitusosuudeksi voidaan katsoa Kansalliskirjaston osallistuminen projektin mm. tarjoamalla palvelintila Fenno-Ugrica -kokoelman tiedostoille ja otta-malla vastuun niiden pitkäaikaissäilytyksestä.

Keskeisimmät muutokset budjetoinnissa koskivat tekijänoikeuksienkorvauksia ja koulutuksia ja ostopalveluita. Korvauksia tekijänoikeuksista ei pilottiprojektin puitteissa maksettu, eikä suunniteltua koulutusta venäläisille kollegoille koskaan järjestetty. Ostopalveluja kertyi budjetoitua enemmän, mutta aineistoja digitoitiin enemmän kuin alun perin oli tarkoituksena, sillä marinkielisten rinnakkaisnimekkeiden määrä nosti hieman kustannuksia tekijänoikeuksien selvityksen ja digitoinnin osalta. Lisäksi projektin varoista maksettiin harvinaisen mordvalaisen sanomalehtiaineiston konservoinnista ja digitoinnista – tätä kuluerää ei oltu allokoitu alkuperäiseen budjettiin, mutta koska budjetissa oli joustonvaraa, niin konservoitavat sanomalehdet päätettiin projektin loppuvaiheessa tilata Venäjän kansalliskirjastolta.

	Vuosi 1.	Vuosi 2.	Yhteensä	KOKO PROJEKTI
PROJEKTIN KULUT TOIMINNOITTAIN	Täydentävä rahoitus (anotaan ulkopuoliselta taholta)	Täydentävä rahoitus (anotaan ulkopuoliselta taholta)	Täydentävä rahoitus (anotaan ulkopuoliselta taholta)	Perusmääräraha
LASKUTUS VENÄJÄLTÄ	**39150**	**114600**	**153750**	**153750**
Digitointi	23000	100000	123000	123000
Tekijänoikeuskorvaukset ja -selvitystyö	10000	14600	24600	24600
Sopimukset ja lisenssit	6150		6150	6150
LAITTEISTO JA JÄRJESTELMÄ	**0**	**20000**	**20000**	**20000**
Julkaisujärjestelmä - ylläpito	0	13000	13000	13000
Julkaisujärjestelmä - kehittäminen	0	7000	7000	7000
KANSALLISKIRJASTON PALKKAKUSTANNUKSI	**30706,25**	**62093,75**	**92800**	**92800**
Tietojärjestelmäasiantuntija 10/2012-03/2013	15800	15800	31600	31600
Kuvailu- ja formaattiasiantuntija	0	21450	21450	21450
Projektipäällikkö 07/2012-10/2013	14906,25	24843,75	39750	39750
MUUT KULUT JA PALKKIOT	**6000**	**37050**	**43050**	**43050**
Verkkosivut	0	2460	2460	2460
Käännöspalkkiot	1000	3920	4920	4920
Seminaarit	0	2460	2460	2460
Matkat ja päivärahat	5000	15910	20910	20910
Koulutukset	0	12300	12300	12300
YHTEENSÄ	**75856,25**	**233743,75**	**309600**	**309600**
Tämä summa muodostuu projektien kuluista ilman yleiskustannuksia, 15%				
	306900 + 46440 = 356 040 euroa			

Kuva 4. *Budjettisuunnitelma toiminnoittain, kesä 2012*

Budjetin vuositoteuma ei vastannut täysin kesällä 2012 tehtyä budjettisuunnitelmaa, ks. kuva yllä. Keskeisimmät syyt tähän olivat sekä rekrytoinnin että ostopalvelujen viivästyminen. Helsingin yliopiston asettaman rekrytointikiellon vuoksi hankkeeseen palkattu tietojärjestelmäasiantuntija pystyi aloittamaan työnsä vasta joulukuussa 2013, jolloin palkkakustannuksia kertyi vuoden 2012 osalta suunniteltua vähemmän. Lisäksi Venäjältä hankittujen palveluiden (digitointi ja tekijänoikeudet) maksut olivat vuoden 2012 osalta alhaisemmat kuin projektia aloittaessa oli allokoitu.

WBS		Kustannuslaji	EUR
4702964	Koneen Säätiö/Suom-Ugrilaisen aineistoje	Tuotot	-356 040,00
4702964	Koneen Säätiö/Suom-Ugrilaisen aineistoje	Kulut	356 141,61
4702964	Koneen Säätiö/Suom-Ugrilaisen aineistoje	Palkat yhteensä	97 442,46
4702964	Koneen Säätiö/Suom-Ugrilaisen aineistoje	Henkilösivukulut	21 990,13
4702964	Koneen Säätiö/Suom-Ugrilaisen aineistoje	Tilakustannukset	
4702964	Koneen Säätiö/Suom-Ugrilaisen aineistoje	Aineet ja tarvikkeet	1 854,59
4702964	Koneen Säätiö/Suom-Ugrilaisen aineistoje	Koneet ja laitteet	
4702964	Koneen Säätiö/Suom-Ugrilaisen aineistoje	Ostetut palvelut	173 292,14
4702964	Koneen Säätiö/Suom-Ugrilaisen aineistoje	Matkat	9 529,81
4702964	Koneen Säätiö/Suom-Ugrilaisen aineistoje	Muut kulut	
4702964	Koneen Säätiö/Suom-Ugrilaisen aineistoje	Yleiskustannukset	52 032,48
Kokonaistulos			101,61

Kuva 5. *Projektin toteuma, SAP-raportti vuoden 2013 osalta (per 31.1.2014)*

9 Tulokset ja yhteenveto

Sukukielten digitointiprojektin pilottihankkeen lähtökohdat olivat yhtä aikaa sekä kiinnostavat että ongelmalliset. Erityisen kiinnostavaksi aloitteen teki se, ettei koskaan aikaisemmin oltu pyritty saattamaan entisen Neuvostoliiton alueella painettuja aineistoja yhtä laajasti avoimeen kansalaiskäyttöön. Ongelmalliseksi aloitteen teki se, ettei sen paremmin Kansalliskirjastolla kuin kenelläkään muullakaan toimijalla ollut aikaisempaa kokemusta tämänluontoisen hankkeen toteuttamisesta ja sen reunaehdoista. Keskeiseksi haasteeksi koettiin ennen kaikkea aineistoihin koskevat tekijänoikeudet ja niiden selvittäminen, sillä aikaisempaa kokemusta Venäjän ja Suomen lainsäädännön soveltamisesta tällaisessa tapauksessa ei ollut. (ks. myös luku 4. Pilottivaiheen riskianalyysi). Haasteista huolimatta, hanke onnistui vähintäänkin kiitettävästi, ennen kaikkea hyvän valmistelevan suunnittelun ja osapuolten välisen yhteistyön ansiosta.

9.1 Asetettujen tavoitteiden saavuttaminen

Mikäli hankkeen onnistumista ja sen tuloksia tarkastellaan sille asetettujen tavoitteiden kautta, niin hanke saavutti sille asetetut odotukset lähes kokonaisuudessaan. Pilotoinnin tarkoituksena oli selvittää

- edellytyksiä Venäjän Kansalliskirjaston kokoelmissa olevien suomalais-ugrilaisia aineistojen digi-taalisesta käyttöön saattamista,
- niihin kohdistuvien tekijänoikeuksien selvittämistä, ja
- aineistojen soveltuvuutta tutkimuskäytössä.

Näistä ylätason tavoitteista keskeisimmät, eli aineistojen käyttöön saattaminen ja tekijänoikeuksien selvitys onnistuivat erinomaisesti. Ainoastaan viimeisen tavoitteen osalta voidaan keskustella kriittisesti, missä suhteessa tavoite on täyttynyt, sillä projektin päättyessä tutkijat eivät ole voineet hyödyntää OCR-editoria omassa tutkimuksessaan. OCR-editorin hidastunutta suunnittelua ja sen valmistumista tuskin olisi voitu näissä olosuhteissa vauhdittaa, sillä pilottihankkeen käytössä olleet resurssit eivät mahdollistaneet kehitystyön jakamista useamman työntekijän kesken. Jatkohankkeen osalta tällaisen työnjakoon on kuitenkin varauduttu. Tämä esiintuotu puute ei kuitenkaan kiistä sitä seikkaa, että pilotointi oli kokonaisuudessaan erityisen onnistunut.

9.2 Pilotoinnin aikana opittua

Mikäli hankkeen onnistumista tarkastellaan myös muutoin kuin yllämainittujen tavoitteiden valossa, niin pilottihankkeen voidaan sanoa ylittäneen sille asetetut tavoitteet. Tätä argumenttia tukevat ainakin seuraavat huomiot:

- tutkijayhteistyö kielentutkimuksen yhteisön kanssa koti- ja ulkomailla kasvoi hankkeen edetessä
- yhteistyö venäläisten kirjastojen kanssa muuttaa tekemisen standardeja pidemmällä aikavälillä
- kansainvälinen tekijänoikeusyhteistyö sai uuden ulottuvuuden
- Kansalliskirjaston asema luotettavana toimijana vahvistui venäläisten partnerien keskuudessa
- oman henkilökunnan ammatillista osaamista voitiin hyödyntää ja lisätä pilotoinnin aikana
- budjettikuri toteutui IT-projekteille harvinaisen hyvin – sitä ei ylitetty.

Hankkeen aikana on harjoitettu tiivistä tutkijayhteistyötä Kieliohjelman turvin työskentelevien kielentutkijoiden kanssa. Tämä yhteistyö on ollut hedelmällistä myös Kansalliskirjastoa ajatel-len, joka ei ole aikaisemmin toiminut yhtä tiiviissä yhteistyössä kielitieteen edustajien kanssa. Ottaen huomioon, että Kansalliskirjasto on tällä saralla uudehko toimija, niin tiivis yhteistyö tutkijoiden kanssa on luonut vankkoja edellytyksiä yhteistyön jatkamiselle. Myös sellaiset kie-lentutkimuksen tieteelliset verkostot, jotka ovat olleet aikaisemmin Kansalliskirjastolle vieraita, ovat tulleet hankkeen aikana tutuimmiksi. Sukukielten digitointiprojektin pilottihankkeen kautta kotimainen tutkijakenttä on alkanut mieltää Kansalliskirjaston luontevaksi yhteistyökumppaniksi. Hyvänä esimerkkinä tästä tendenssistä voidaan mainita toisaalta OCR-editorin suunnitteleminen, mutta myös osallistuminen digital humanities -kentällä tapahtuvaan keskusteluun. Jatkohankkeen toteutuessa yhteistyötä tiivistetään edelleen osallistumalla yhdessä tieteelliseen keskusteluun, konferensseihin, kansalaistieteiden järjestämiseen.

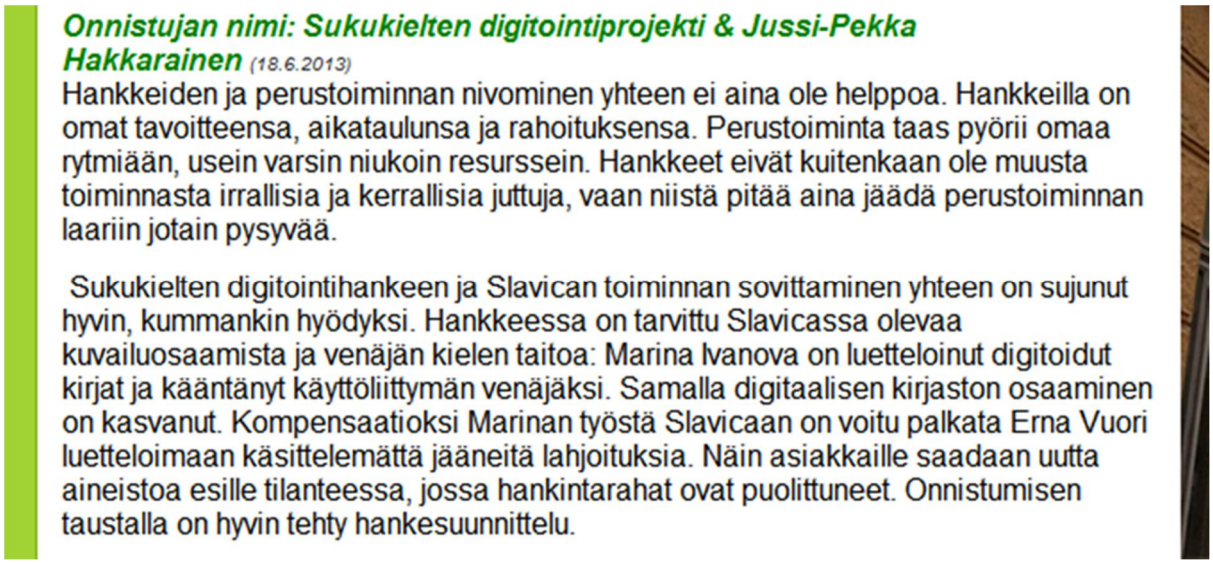
Kansalliskirjastolla on perinteisesti ollut vahva kokemus yhteistyöstä venäläisten kirjastojen kanssa. Viime vuosina aikana yhteistyö on ollut tiiveintä ennen kaikkea Venäjän kansalliskir-jaston kanssa, mutta myös Karjalan tasavallan kansalliskirjasto on ollut keskeinen kirjastoalan yhteistyökumppani. Ennestään vahvat suhteet ovat saaneet Sukukielten digitointiprojektissa konkreettisia muotoja ja tuloksia, jotka ovat olleet kehittämässä venäläisten kirjastojen toimin-taa tuleville vuosille. Ne toimintatavat ja -käytänteet, joita pilotoinnin aikana voineet venäläisille kirjastoille esitellä ja perustella ovat osaltaan avaamassa venäläisissä kirjastoissa olevia digitoituja aineistoja laajemmalle käyttäjäkunnalle. Olemme pilotoinnin aikana puhuneet venäläisille kirjastoille aineiston avoimesta saavutettavuudesta ja suositelleet useissa tapauksissa digitointien saattamista avoimeen verkkokäyttöön. Puutteita on myös standardisoinnissa, sillä digitoidulle aineistolle ei ole useinkaan asetettu yhteisiä

laatumäärityksiä, mikä on ollut esteenä aineiston hyödyntämiselle mm. kielentutkimuksen parissa. Samassa yhteydessä on pyritty myös lobbaamaan OCR-tekstin tuomista osaksi digitointiprosesseja, mikä ei aina ole ollut käytäntönä edes suurissa valtiollisissa kirjastoissa. Tässä suhteessa perustelut ja keskustelut aineistojen avoimesta verkkokäytöstä ovat mahdollistaneet asenneilmaston muutosta. Kansalliskirjastolla ja Sukukielten digitointiprojektilla on ollut selkä rooli eri kokonaisuuksia yhdistävänä organisaationa, jonka osaaminen digitointiin, digitaalisten aineistojen esittämiseen ja kokemus kansainvälisestä yhteistyöstä avaa maailmaa myös venäläisille toimijoille.

Asenneilmasto on muuttunut myös tekijänoikeudenselvityksen myötä. Sukukielten digitointi-hankkeessa tehty tekijänoikeuksienselvitys tehtiin tässä muodossaan ensi kertaa. Venäläisissä kirjastoissa on vallinnut skeptisyyttä tekijänoikeuksien selvittämistä kohtaan ylipäänsä, mutta hankkeessa voitiin osoittaa, että harjoitettu menettelytapa on mahdollinen. Pilottihankkeen aikana ilmeni, että tekijänoikeuksienselvityksen suorittanut järjestö, National Library Resource, saanee laissa määritellyn aseman venäläisen kirjastoalan lisenssiorganisaationa, mutta Venäjän duuma ei ole vielä päättänyt lakimuutoksesta. On siis mahdollista, että jatkossa kirja-alan tekijänoikeuksia selvitetään laajemminkin Sukukielten digitointiprojektissa suoritetulla tavalla. (Ks. myös National Library Resourcen johtajan kirjoitus selvityksestä hankkeen blogissa: https://blogs.helsinki.fi/fennougrica/2013/05/20/gordian-knot-of-copyrights/)

Tekijänoikeuksien selvityksen ja asenneilmaston tuulettamisen seurauksena Kansalliskirjaston asema luotettavana ja vastuullisena toimijana venäläisten kirjastojen silmissä on vahvistunut pilotoinnin aikana. Tämä saavutettu maine ja luottamus on hyödyksi myös tulevaisuuden projektin kannalta, sillä hankkeessa on luotu edellytyksiä laajemmallekin kirjastoalan yhteistyölle.

Projektin aikana on pyritty kasvattamaan myös Kansalliskirjaston omaa osaamista. Periaatteena on ollut, ettei hanke kuluttaisi Kansalliskirjaston resursseja, vaan myös kirjasto hyötyisi osaamisen kasvusta, eli osaavan henkilökunnan hyödyntäminen hankkeessa täytyy tuottaa jotain uutta Kansalliskirjastollekin. Pilottivaiheessa käytettiin hyväksi ennen kaikkea Slaavilaisen kirjaston (Slavica) ja kirjastoverkkopalveluiden osaamisvarantoja. Vakituisen henkilökunnan työpanosta siirrettiin Sukukielten digitointiprojektin myönnöltä työpanosta vastaava korvaus niin kirjastoverkkopalveluille kuin Slaavilaiselle kirjastollekin. Slaavilaisessa kirjastossa kompensaatiorahalla työskenteli ylimääräinen kirjastosihteeri ja kirjastoverkkopalveluissa voitiin palkata yksi tietojärjestelmäasiantuntija lähes puoleksi vuodeksi.



Kuva 6. *Kansalliskirjaston intranet.*

9.3 Tulokset suhteessa Koneen Säätiön Kieliohjelmaan

Pilotoinnin onnistumista on syytä arvioida myös Koneen Säätiön Kieliohjelmassa esitettyjen tavoitteiden ja pyrkimysten valossa. Koneen Säätiön kieliohjelman päätavoitteena on edistää pienten suomalais-ugrilaisten kielten, suomen sekä Suomen vähemmistökielten dokumentoin-tia. Lisäksi aineistojen saatavuuden, saavutettavuuden ja käytettävyyden parantamisen osalta Kieliohjelman tavoitteena on saattaa sekä vanhoja että uusia kieliaineistoja tiedeyhteisön ja muun yhteiskunnan avoimeen käyttöön. Erikseen Kieliohjelmassa mainitaan, että ”Pietarissa sijaitsevassa Venäjän kansalliskirjastossa on runsaasti suomalais-ugrilaisilla pienillä kielillä kirjoitettuja aineistoja, kuten Neuvostoliiton alkuvuosina tehtyjä oppikirjoja, valistuskirjallisuutta, kaunokirjallisuutta, lehtiä ja aikakauskirjoja. Kieliohjelman tavoitteena on saada esimerkiksi näitä aineistoja avoimeen käyttöön.”

Sukukielten digitointiprojekti on pystynyt vastaamaan näihin Kieliohjelmassa esitettyihin tavoitteisiin ja haasteisiin kiitettävällä tavalla. Pilottihankkeen aikana digitoitu aineisto on saatettu avoimeen kansalaiskäyttöön Fenno-Ugrica -kokoelmassa ja näin on tuotettu uusia kieliaineistoja sekä tutkijayhteisön, että yhteiskunnan avoimeen käyttöön.

Saavutettavuutta ja käytettävyyttä on pyritty käyttöön asettamisen yhteydessä huomioimaan mahdollisimman laajasti. Fenno-Ugrican käyttöliittymä on käännetty myös venäjäksi ja julkaisuarkistoa on pyritty optimoimaan venäläisten hakukoneita varten, jolloin aineiston käytettä-vyys ja saavutettavuus on parantunut. Lisäksi aineiston kuvailuun, sen metadataan, on kiinnitetty erityistä huomiota.

Sukukielten digitointiprojektissa on pyritty vastaamaan myös sirpaleisen tiedon yhteen saattamiseen ja auttamaan niin tutkijoita kuin muitakin käyttäjiä löytämään sopivia kieliaineistoja. Viime aikoina suomalais-ugrilaisten aineistojen digitointi

Venäjän federaation alueella toimivissa kansalliskirjastoissa on kasvanut merkittävästi viimeisten vuosien aikana. Monet venäläi-set kirjastot ovat muodostaneet omia digitaalisten aineistojen kokoelmiaan ja antaneet yleisölle joko avoimen tai rajatun mahdollisuuden käyttää aineistoa. On ajateltu, että sekä uusien että vanhojen suomalais-ugrilaisten aineistojen digitointi ja aineistojen käyttöön saattaminen tukee sukukielten puhujien kielen oppimista, sillä näiden kielten voidaan ajatella edustavan omien yhteisöjensä, kulttuuriensa ja perinteidensä tärkeää voimavaraa ja maailmankuvaa, ja kulttuuri, perinteet ja maailmankatsomus välittyvät parhaiten äidinkielisen vapaassa kielenkäytössä.

Samaan aikaan kuitenkin tiedonjakaminen ja -harmonisointi ovat jääneet digitointiinnon jalkoihin ja venäläisten kansalliskirjastojen kanssa käytyjen keskustelujen myötä on tullut selväksi, ettei tällä hetkellä kukaan kontrolloi kuinka paljon ja mitä suomalais-ugrilaisia aineistoja Venäjän federaation alueella digitoidaan – tämä vaikeuttaa aineistojen saavutettavuutta, digitointiprojektien järjestelmällistä suunnittelua ja voi johtaa päällekkäiseen työhön, jolloin resursseja käytetty tehottomasti.

Samaa ongelmaa kuvastaa aineiston sirpaleinen saavutettavuus. Loppukäyttäjien näkökulmasta monien käyttöliittymien käyttö ja niiden hallinta vaatii aikaa ja tietoteknistä osaamista, mutta myös valtakielen hallitsemista. Ei ole olemassa järjestelmää, minkä avulla kaikkia digitoituja suomalais-ugrilaisia aineistoja voitaisiin etsiä, selata ja tarvittaessa jopa käyttää.

Edistääkseen Koneen Säätiön Kieliohjelmassa kuvattuja pyrkimyksiä Kansalliskirjasto toteutti Fenno-Ugrica -kokoelman lisäksi Opetus- ja kulttuuriministeriöltä saadulla avustuksella yhteistyöhankkeen, jonka tavoitteena oli luoda yhteinen ja avoin tietojärjestelmäinfrastruktuuri eri kirjastoissa digitoiduille suomalais-ugrilaisille kieliaineistoille. Uralica-portaali avattiin elokuussa 2013 ja siihen linkitetty digitoituja aineistoja Karjalan tasavallan kansalliskirjaston (Petroskoi) ja Udmurtian tasavallan kansalliskirjaston (Iževsk), Göttingenin valtion ja yliopiston kirjaston (Niedersächsisches Staats- und Universitätsbibliothek) ja Viron kielen instituutin (Eesti Keele Instituut, Tallinna) kokoelmista.

10 Suositukset jatkohankkeen toteuttamista varten

Seuraavassa esitellään sellaisia suorituksia mahdollisen jatkohankkeen toteuttamista ajatellen, joiden huomioonottaminen olisi tarkoituksenmukaista hankkeen hyvälle toteuttamiselle.

Jatkohankkeen aikana on tarkoitus digitoida ja saattaa käyttöön lähes 1100 monografia- ja 51 sanomalehtinimekettä. Monografiasivuja suunnitelman mukaisesti kertyy noin 88 300 ja sano-malehtisivuja noin 72 500. Digitoitavat aineistot on valikoitu yhteistyössä kotimaisen tutkija-kunnan kanssa ja sen on katsottu palvelevan niin kotimaista kuin ulkomaista fennougristiikan alan tutkimusta. Hankkeen toteutuessa aineistot muodostaisivat maailman suurimman uralilaisten kielten resurssin, joka toisi tutkijakunnan ulottuville sellaisia kieliaineistoja, joihin aikaisemmin heillä ei ole ollut mahdollista tutustua ja johon kaikilla käyttäjillä on asuinpaikasta riippumatta avoin pääsy. Aineisto saatetaan sekä tutkijoiden hyödynnettäväksi että avoimeen kansalaiskäyttöön Kansalliskirjaston ylläpitämässä Fenno-Ugrica –kokoelmassa (fennougrica.kansalliskirjasto.fi)

Jatkohankkeen aineistojen kohdalla on pyritty painottamaan ennen kaikkea hyvin vähän digitoituja sekä vaikeasti saavutettavia alueellisia ja periferisiä julkaisuja. Sisällön puolesta keskuksen ulkopuolella sijaitsevien alueiden kieli (paikallislehdet) on keskuksen kieltä mielenkiintoisempaa, koska periferiassa voi ilmetä joko keskuksen kielestä poikkeavaa murrevariaatiota tai konservatiivisuutta myös kirjoitetussa kielessä. Paikallisten sanomalehtien digitointia puoltaa myös pyrkimys edistää aineiston saavutettavuutta: paikallismateriaaliin keskittymällä tuodaan tutkijoiden ulottuville aikaisemmin huonosti saatavaa materiaalia ja digitoida sellaisia sanomalehtiä, joiden asema venäläisten kirjastojen digitointisuunnitelmissa on joko hyvin marginaalinen tai jopa poissaoleva.

10.1 Toteutustapa

Jatkohankkeen osalta toteutustapa noudattelee pääsääntöisesti pilottihankkeen aikana luotuja ja testattuja käytänteitä muutamin pienin poikkeuksin. Aineisto digitoidaan pääsääntöisesti Venäjän Kansalliskirjaston (Pietari) kokoelmista ja aineistoja koskevia tekijänoikeuksia selvitetään yhteistyössä National Library Resource (Moskova) kanssa. Aineisto luovutetaan vain TIFF-muotoisena Kansalliskirjastolle jatkokäsittelyä varten. Jatkokäsittely suoritetaan joko Kansalliskirjaston Digitointi-, mikrofilmaus- ja konservointiyksikössä Mikkelissä tai yhteistyössä tutkijoiden kans-

sa. Tavoitteena on rakenteellistaa kuvamuotoista aineistoa, eli siihen lisätään OCR-tunnistus ja palstoitus. Rakenteellistettu aineisto siirretään PDF-muotoisena osaksi digitaalista Fenno-Ugrica -kokoelmaa, missä aineisto myös kuvaillaan. Tutkijat voivat hyödyntää palstoitettua aineistoa rikastaessaan materiaalia OCR-editorin avulla. OCR-editorilla korjattu kielimateriaali luovutetaan edelleen FIN-CLARINin ylläpitämään Kielipankkiin ja se on siten myös muun tutkijayhteisön käytettävissä.

Sukukielten digitointiprojektia johtaisi Suomen Kansalliskirjasto ja se vastaa yhteistyöstä ja sen koordinoinnista kansainvälisten ja kotimaisten toimijoiden kanssa. Keskeisimpiä ulkomaisia toimijoita ovat Venäjän Kansalliskirjasto Pietarissa, jonka kokoelmista suurin osa Sukukielten digitointiprojektin aineistoista digitoidaan. Koska jo aineistojen digitointiin vaaditaan tekijänoikeuksien haltijoiden lupa, niin digitoitavien aineistojen tekijänoikeuksia selvittää molempien kansalliskirjastojen toimeksiannosta moskovalainen tekijänoikeusjärjestö National Library Resource. Toimijoiden välinen työnjako noudattelee pilottivaiheen aikaisia käytäntöjä ja määrittelee toimijoiden vastuualueita: Venäjällä selvitetään paikalliset tekijänoikeudet ja tutkijoiden valitsemat aineistot digitoidaan venäläisistä kokoelmista, mutta aineisto saatetaan käyttöön Suomessa. Tämä tuotantomalli on kansainvälisesti ainutlaatuinen ja avaa uusia yhteistyömahdollisuuksia Venäjän ja lännen välillä niin kirjastojen välisellä kuin humanistisen tutkimuksen kentällä yleensä.

10.2 Jälkikäsittely

Yksi jatkohankkeen tavoitteista koskee Venäjän federaation alueella tehtävää digitointia ja sen laatua. Yhteistyön avulla suomalais-ugrilaisten aineistojen digitoinnille voitaisiin luoda yhdessä määritellyt laatukriteerit, sillä yhtenäinen, standardinmukainen digitointijälki ja kuvailutieto mahdollistavat myös aineiston pitkäaikaissäilytyksen ja turvaa sen säilyvyyden tuleville sukupolville.

Sukukielten digitointiprojektilla on kiinnekohtia myös kieliteknologiseen tutkimukseen, sillä projektin yhdeksi päämääräksi voidaan laajasti ajatella digitaalisten kirjasto- ja arkistoaineisto-jen käyttötapojen ja käytettävyyden parantaminen. Projektissa edistetään suomalais-ugrilaisen aineiston käyttöön saattamisen lisäksi menetelmiä, joilla digitoitua raakadataa voidaan jalostaa entistä käyttökelpoisemmiksi aineistoiksi ja joilla tällaista aineistoa voidaan hyödyntää. Näillä menetelmillä tarkoitetaan digitoidun aineiston OCR-tunnistuksen lisäämistä, tunnistetun tekstimassan palstoittamista sekä ennen kaikkea kielenkorjaukseen tarkoitetun OCR-editorin jatkokehittämistä, jonka avulla voidaan korjata digitoinnin ja OCR-tunnistuksen yhteydessä jääneitä virheitä. Pilottivaiheen aikana kehittämä OCR-editorin edelleen kehittäminen tukisi siis myös digitointialan kokonaisvaltaista kehittämistä

Projektin pilottivaiheen kuluessa käytettiin hyvin paljon työtä Pietarista saatujen PDF-dokumenttien analysoinnissa ja ALTO XML-tiedostojen tuottamisessa. Jatkossa on pyrittävä saamaan digitoitu aineisto mieluiten digitoijan tuottamana ALTO XML:nä. Formaatti on digi-toinnissa käytettävien ohjelmistojen yleisesti tukema.

Pilottiprojekti on ollut verraten erillinen osa Kansalliskirjaston kehitystoiminnassa. Siinä ei ole pystytty täysin käyttämään hyväksi kirjaston asiantuntijoita esim. käytettävyys- ja graafisen suunnittelun, laitteistopalvelujen alalla - ja yleensä tukena ja ajatusten kaikupohjana suunnittelun ja toteutuksen aikana. Projektin jatkuessa työn järjestelyyn ja johtamiseen on näiltä osin löydettävä toisenlaisia ratkaisuja. Työssä voisi myös paremmin käyttää hyväksi Kansalliskirjaston tehtävienhallinnan (JIRA) ja muita välineitä.

OCR-editori lienee tällaisenaan ainoa lajissaan, mutta tähän mennessä siihen kehitetyt toimin-nallisuudet eivät vastaa vielä tutkijakunnan tarpeita kokonaisuudessaan, joten jatkohankkeessa tulisi tätä työkalua edelleen kehittää yhteistyössä niin tutkijoiden kuin eri hankkeiden kanssa jatkohankkeen aikana. OCR-editorin kehitystyö tehtäisiin Kansalliskirjaston kirjastoverkkopalveluissa ja kehitystyöhön tulisi palata jatkohankkeen varoista toinen tietojärjestelmäasiantuntija, jonka työtehtävät liittyisivät editorin jatkokehittämiseen yhdessä jo käytettävissä olevan resurssin kanssa.

10.3 Tutkijayhteistyö ja kansalaistieteet

Jatkohankkeessa digitoitavaksi esitetty aineisto on valittu yhdessä tutkijakunnan kanssa. Tuot-tamalla tämän aineiston Kansalliskirjasto vahvistaa omaa tutkijayhteistyötään kielentutkimuk-sen partnerina, mutta vahvistaa myös yhteistyötä venäläisen kirjastokentän kanssa. Sukukielten digitointiprojektin jatkohankkeen keskeisiä partnereita ovat erityisesti Koneen Säätiön Kieliohjelman apurahalla työskentelevät tutkijat ja tutkijayhteisöt. Jatkohankkeen kannalta keskeisessä asemassa ovat seuraavat tutkimushankkeet ja -projektit:

- Turun yliopisto, Volgan alueen kielten tutkimusyksikkö: ”Marin kielen sanaston kehitys 1920- ja 1930-luvuilla”
- Helsingin yliopisto, Jack Rueter ja työryhmä: ” Morfologisten jäsentimien luominen suomalais-ugrilaisille vähemmistökielille”
- Helsingin yliopisto, Riho Grünthal ja työryhmä: ”Itämerensuomalaisten kielten muutos monikielisessä ympäristössä”
- Helsingin yliopisto, Suomalais-ugrilaiset kielet ja internet -hanke.
- Kotimaisten kielten tutkimuskeskuksen vanhan kirjasuomen sanakirjat
- Kirjallisuuspankki-hanke
- FIN-CLARIN
- HFST (Helsinki Finite-State Transducer Technology)
- Tromssan yliopisto, Giellatekno

Kotimaisen tiedeyhteistyön on tarkoitus olla pelkkää aineiston tuottamista laajempaa ja ylittää tieteiden välisiä rajoja. Ylirajaista yhteistyötä jatketaan jatkohankkeessa OCR-editorin tiimoilta, jota on kehitetty pilottihankkeessa vuoden 2012 lopulta lähtien. Editoria hyödynnetään myös eräissä muissa Koneen Säätiön rahoittamissa hankkeissa, kuten Kotimaisten kielten tutkimuskeskuksen vanhan kirjasuomen sanakirjan 3. osan loppuunsaattamiseen tähtäävässä hankkeessa. Tarkoituksena on, että OCR-korjattu vanhan kirjasuomen aineisto liitetään Kotuksen tekstikorpuskokoelmaan ja tuodaan muiden korpusten tavoin julkisesti saataville.

Kirjallisuuspankki on Suomen kielen, suomalais-ugrilaisten ja pohjoismaisten kielten ja kirjallisuuksien laitoksen, Kansalliskirjaston ja Helsingin yliopiston kirjaston yhteishanke. Se on kattava ja lähdekriittinen kokonaisuus suomen- ja ruotsinkielistä kaunokirjallisuutta sekä siihen liittyvää tutkimustietoa, kulttuurihistoriallinen resurssi koko maan tutkimukselle, koulutukselle ja kansansivistystyölle. Kirjallisuuspankin aineisto on niin ikään kaikkien vapaasti käytettävissä internetin välityksellä. Kirjallisuuspankissa julkaistavien digitoitujen klassikkotekstien stilisointi on tarkoitus suorittaa OCR-editoria hyödyntäen.

Pilottihankkeessa on rakennettu kansainvälistä yhteistyöverkostoa ja jatkohankkeen aikana tulisi yhteistyötä syventää ainakin seuraavien toimijoiden kanssa projektin eri osa-alueilla:

- Venäjän kansalliskirjasto, Pietari
- Venäjän Valtiollinen kirjasto, Moskova
- National Library Resource, Moskova
- Udmurtian tasavallan kansalliskirjasto, Iževsk
- Karjalan tasavallan kansalliskirjasto, Petroskoi
- Viron kielen instituutti, Tallinna
- Tarton yliopisto, Tartto
- Valtion- ja yliopistonkirjasto, Göttingen
- Max Planck-Institut für Wissenschaftsgeschichte, Berliini
- Suomen Moskovan-suurlähetystö
- Hanti-Mansian tasavallan kansalliskirjasto, Hanti-Mansijsk
- Komin tasavallan kansalliskirjasto, Syktyvkar
- Marin tasavallan kansalliskirjasto, Joškar-Ola
- Mordovian tasavallan kansalliskirjasto, Saransk

Sukukielten digitointiprojektin jatkohankkeen yksi keskeinen tavoite on luoda kansalaistieteiden verkosto Fenno-Ugricassa olevien aineistojen ja jatkohankkeen kumppanuuksien ympärille. Tarkoituksena on kerätä kansalaistieteilijöiden joukko, joka pystyvät toimimaan paikallisissa olosuhteissa talkoistamisen ohjaajina niin oppilaitoksissa kuin vaikkapa alueellisissa kirjastoissa. Tällä tavoin Sukukielten

digitointiprojekti pyrkisi jatkohankkeen aikana talkoistamaan Fenno-Ugricassa olevan aineiston käsitellen sitä OCR-editorilla.

Sopivien kansalaistieteilijöiden kouluttamista Venäjän federaation alueella pyritään mahdollis-tamaan Suomalais-venäläisen kulttuurifoorumin (http://www.kultforum.org) kautta. Vuonna 2014 suomalaiset partnerit saavat ehdottaa yhteistyöhankkeita ja Kansalliskirjasto pyrkii mah-dollistamaan kansalaistieteiden levittämisen luomalla tarvittavat partneriverkostot niin paikallisten kirjastojen kuin koulujenkin kanssa. Sopivien partnerien paikantamisessa työskennellään yhteistyössä tutkijakunnan kanssa.

11 Yhteystiedot ja tiedustelut

Liisa Savolainen
johtaja
Kansalliskirjasto
Tutkimuskirjasto
PL 15 (Fabianinkatu 35)
00014 HELSINGIN YLIOPISTO
liisa.savolainen@helsinki.fi
02941 22745

http://www.kansalliskirjasto.fi/kokoelmatjapalvelut/digitaalisetkokoelmat/sukukielet.html